

**APPLICATION FOR PATENT**  
**FOR**  
***BACILLUS THURINGIENSIS* CHROMOSOMAL GENOME SEQUENCES AND USES**  
**THEREOF**

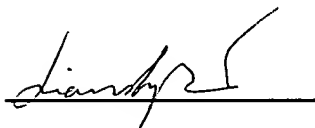
by

**David R. Corbin, Thomas M. Malvar, and Hridayabhiranjan Shukla**

I hereby verify the above-identified patent application is being deposited with  
the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10  
Express Mail Label No. EK016827222US on the date indicated and is addressed to:  
Commissioner for Patents, Washington, D.C. 20231

**Jian S. Zhou**

Registration No. 41, 422



Date: Sept. 15, 2000

## CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority under 35 U.S.C §119(e) of U.S. Provisional Application Serial No. 60/154,678 filed on September 17, 1999, the entire content of which is incorporated herein by reference.

5

## FIELD OF THE INVENTION

The present invention relates to nucleic acid sequences from *Bacillus thuringiensis* and, in particular, to chromosomal genomic DNA sequences. The invention encompasses nucleic acid molecules present in non-coding regions as well as nucleic acid molecules that encode proteins and fragments of proteins. Nucleic acid sequences that encode proteins and/or enzymes and homologues and fragments thereof are encompassed by the invention including but not limited to insect inhibitory proteins, proteins capable of conferring antibiotic resistance, microbial inhibitory proteins including bactericidal, bacteriostatic, fungicidal, and fungistatic proteins, polyketide synthases, transposons and mobile genetic elements and their corresponding transposases, excisases and integrases, phage and phage particle proteins, other useful protein homologues, ribosomal RNA (rRNA), and transfer RNA (tRNA). In addition, proteins and fragments thereof so encoded and antibodies capable of binding the proteins are encompassed by the present invention. The invention also relates to methods of using the disclosed nucleic acid molecules, proteins, fragments of proteins, and antibodies, for example, for gene identification and analysis, preparation of constructs, transformation of cells with nucleotide compositions disclosed herein to produce *Bacillus thuringiensis* proteins or fragments thereof, in particular novel insect inhibitory, bactericidal, fungicidal and nematocidal proteins.

## BACKGROUND OF THE INVENTION

*Bacillus thuringiensis* is a spore-forming Gram-positive bacterium. During sporulation, *B. thuringiensis* produces proteinaceous inclusions which are composed of proteins known as insecticidal crystal proteins (ICPs), Cry proteins, or delta-endotoxins. These proteins are toxic to a variety of insect species including orders Lepidoptera, Coleoptera, Diptera, Hemiptera, Hymenoptera, Orthoptera, and Mallophaga and to nematodes, mites, and protozoa (Beegle and Yamamoto, *Can. Entomol.* 124:587-616; Feitelson, *Advanced Engineered Pesticides* (L. Kim, ed.), Marcel Dekker, Inc., New York (1993), pp. 63-71; Feitelson, *et al.*, *Bio/Technology* 10:271-275; U.S. Patent No. 4,948,734 (1990)). Due to their high specificity for particular insect pests and their safety for man and the environment, ICPs have been used as biopesticides for the last three decades. Using molecular genetic techniques, numerous delta-endotoxin

genes have been isolated and their DNA sequences determined. The cloning and sequencing of a number of  $\delta$ -endotoxin genes from a variety of *B. thuringiensis* strains has been described and are summarized by Schnepf *et al.* (*Microbiol. Mol. Biol. Rev.* 62:775-806, *Bacillus thuringiensis* And Its Pesticidal Crystal Proteins, 1998). The nomenclature and appearance of newly identified genes is summarized and regularly updated at [http://www.biols.susx.ac.uk/Home/Neil\\_Crickmore/Bt/](http://www.biols.susx.ac.uk/Home/Neil_Crickmore/Bt/). These genes have been used to develop certain genetically engineered *B. thuringiensis* products that are in commercial use. Recent developments have seen new  $\delta$ -endotoxin delivery systems developed, including genetically engineered plants that contain and express  $\delta$ -endotoxin genes. *Bacillus thuringiensis* is a key source of genes, which when modified can be used for transgenic expression to provide pest resistance in plants.

*B. thuringiensis* strains are classified into subspecies or varieties, based on biochemical and serological criteria (de Barjac, *Entomophaga* 7: 5-61 (1962); de Barjac, *Proceedings of the IIIrd International Colloquium on Invertebrate Pathology* (C.C. Payne and H.D. Burges, eds.), Society for Insect Pathology, U.K., 451-453 (1982)). Each subspecies may produce one or several insecticidal protein toxins. To date, approximately 172  $\delta$ -endotoxins belonging to 28 classes have been identified. There is also a nonprotein toxin, the  $\beta$ -exotoxin, secreted by some *B. thuringiensis* strains. This toxin, which is assayed on house fly larvae (Sêbesta *et al.*, "Thuringiensin, the  $\beta$ -exotoxin of *Bacillus thuringiensis*," in W.H. Burgess (ed.), *Microbial Control of Pests and Plant Diseases, 1970-1980*, Academic Press, Inc., New York, pp. 249-281 (1981)), is not as selective as the  $\delta$ -endotoxins.

Extensive studies have been carried out with *B. thuringiensis* subspecies that produce proteinaceous inclusions during sporulation. The inclusions are often bipyramidal, but some are cuboidal or multifaceted, and there is a wide variety of other morphologies. Some strains contain more than one type of inclusion in each cell. These inclusions are present within the mother cell adjacent to the spore, but in a few subspecies, they are localized within the exosporium (Aronson *et al.*, *Bacteriol. Rev.* 40:360-402 (1976)). Inclusions are released, as is the spore, upon cell lysis.

*Bacillus* strains can have a chromosomal genome size of 2.4 to 5.7 Mbp (Carlson, *et al.*, *Appl. Environ. Microbiol.* 60: 1719-1725 (1994)). Physical maps of chromosomes of two *B. thuringiensis* strains, *B. thuringiensis* subsp. *Berliner* 1715 and *B. thuringiensis* subsp. *Thuringiensis* HD2, have been constructed and are estimated to be between 5.4 and 5.7 Mbp (Carlson, *et al.*, *Microbiol.* 142: 1625-1634 (1996); Carlson and Kolstø, *J. Bacteriol.* 175: 1053-1060 (1993)). The total genomes of each of these two strains consist of one or more chromosomes, and a more variable component comprised of extrachromosomal elements (Carlson and Kolstø, *Mol. Microbiol.* 13:161-169 (1994)).

Most *B. thuringiensis* isolates have several extrachromosomal elements, some of them circular plasmids and others linear (Carlson, *et al.*, *Microbiol.* 60: 1719-1725 (1994)). In general, crystal-protein genes are localized on large plasmids (*ca.* 40 to 200 Mda) of *B. thuringiensis* (Gonzalez, *et al.*, *Plasmid* 5: 351-365 (1981); Carlton and Gonzalez, *Molecular Biology of Microbial Differentiation*, American Society for Microbiology, Washington, D.C. 246-252 (1985), Kronstad, *et al.*, *J Bacteriol.* 154: 419-428 (1983)), and in some cases, more than one gene is present on a given plasmid (Aronson *et al.*, *Bacteriol. Rev.* 40:360-402 (1976); Carlton *et al.*, "The genetics and molecular biology of *Bacillus thuringiensis*," in D.A. Dubnau (ed.), *The Molecular Biology of the Bacilli*, Vol. II, Academic Press, Inc., New York, pp. 211-249 (1985)). However, chromosomal crystal-protein genes have been reported in some *B.*

10 *thuringiensis* strains (Carlson and Kolstø, *J. Bacteriol.* 175: 1053-1060 (1993), Klier, *et al.*, *EMBO J* 1: 791-799 (1982), Kronstad, *et al.*, *J Bacteriol.* 154: 419-428 (1983)).

*Bacillus thuringiensis* strains often contain multiple epigenetic elements which are known to harbor genes expressing vegetative insecticidal proteins (VIP's) and Bt crystalline insecticidal and nematocidal proteins. It is believed that many other Bt insecticidal / nematocidal genes are present within

15 these sequences, some of which may only be expressed under conditions which cannot be artificially simulated, some of which may be cryptic, and some of which may be actively expressed but which have not been previously identified due to their limited availability as a result of very low levels of expression. Identification of whole or substantial portions of DNA sequences of individual plasmids would greatly facilitate identification of genes encoding novel insect inhibitory proteins. However, when one tries to

20 isolate and purify plasmid DNA of a *B. thuringiensis* species for constructing genomic DNA libraries used in sequencing, it would be difficult to eliminate the contamination of chromosomal DNA. Such contamination would complicate greatly the sequencing effort of individual plasmids and subsequently hinder construction of genetic maps of individual plasmids of the *B. thuringiensis* species. Thus, it would be desirable to generate the complete DNA sequence of the chromosomal genome exclusive of epigenetic

25 sequences of a *B. thuringiensis* species, because the complete DNA sequence of the chromosome could be used as a background to significantly minimize the interference of chromosomal DNA sequences in identification of whole or a substantial portion of individual plasmids and of novel genes encoding insect inhibitory proteins.

Furthermore, although it is unexpected that the complete DNA sequence of the *Bacillus*

30 *thuringiensis* chromosomal genome exclusive of epigenetic sequences would provide a substantial number of Bt crystalline insecticidal / nematocidal and VIP genes for second generation insect / pest control in crop species, comparison of the open reading frames present within the *Bacillus thuringiensis* chromosomal genome with other bacterial genome sequences, in particular other *Bacillus* species

genomic sequences would allow the subtraction of common sequences and thus the identification of sequences novel and unique to *Bacillus thuringiensis*, and which may play a role in the regulation of expression or activity of genes encoding insecticidal proteins, and may also provide a plethora of useful genes for future insect resistance management technologies and applications. Therefore, it is advantageous to generate the complete DNA sequence of the chromosomal genome exclusive of epigenetic sequences of a *B. thuringiensis* species.

Chromosomal genome sequence information from *B. thuringiensis* allows comparisons of those sequences with sequences from other *B. thuringiensis* strains as well as comparisons with DNA sequences from other organisms, including plants, mammals such as humans, bacteria, and fungi such as yeasts. In addition, genome sequencing and mapping provides increased opportunities for identification and isolation of agents of commercial interest, as well as insight into mechanisms of genome interactions.

### SUMMARY OF THE INVENTION

The present invention provides an isolated and purified nucleic acid molecule having a first nucleotide sequence, wherein: (1) the first nucleotide sequence hybridizes under stringent conditions to a second nucleotide sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283 or complement thereof, wherein the hybridizing portion of the second nucleotide sequence is at least 50 nucleotides in length; (2) the first nucleotide sequence is a portion of third nucleotide sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283; or (3) the first nucleotide sequence is the complement of (1) or (2).

In a preferred embodiment, the hybridizing portion of the second nucleotide sequence is at least 100 nucleotides in length. In a more preferred embodiment, the hybridizing portion of the second nucleotide sequence is at least 200 nucleotides in length. In a further more preferred embodiment, the hybridizing portion of the second nucleotide sequence encodes any polypeptide or protein or set forth in Table 1.

The present invention also provides an isolated and purified nucleic acid molecule comprising a nucleotide sequence, wherein: (1) the nucleotide sequence encodes any polypeptide or protein set forth in Table 1; or (2) the nucleotide sequence is the complement of (1).

The present invention, in another aspect, provides a substantially purified polypeptide or protein comprising an amino acid sequence, wherein the amino acid sequence is defined as follows: (1) the amino acid sequence is encoded by a first nucleotide sequence which specifically hybridizes to the complement of a second nucleotide sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283; or (2) the amino acid sequence is encoded by a third nucleotide sequence

that is at least 50% identical to a portion of the complement of a sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283. In alternative embodiments, the above described third nucleotide sequence is at least 55% identical, at least 60% identical, at least 65% identical, at least 70% identical, at least 75% identical, at least 80% identical, at least 85% identical, at least 90% identical, or at least 95% identical to a portion of a sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283; and, the above described third nucleotide sequence is identical to a portion of a sequence selected from SEQ ID NO: 1 through SEQ ID NO: 8283.

The present invention also provides a recombinant construct comprising: (A) a promoter region which functions in a host cell to cause the production of an mRNA molecule; which is operably linked to (B) a structural nucleotide sequence, wherein the structural nucleotide sequence encodes a polypeptide or protein set forth in Table 1; which is operably linked to (C) a 3' non-translated sequence that functions in said cell to cause termination of transcription.

The present invention also provides a recombinant construct comprising: (A) a promoter region which functions in a host cell to cause the production of an mRNA molecule wherein the promoter region is selected from the group consisting of promoter sequences located within SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof; which is linked to (B) a structural nucleotide sequence encoding a polypeptide; which is linked to (C) a 3' non-translated sequence that functions in said cell to cause termination of transcription.

The present invention also provides a transformed cell having an exogenous nucleic acid molecule which comprises: (A) a promoter region which functions in said cell to cause the production of an mRNA molecule; which is operably linked to (B) a structural nucleic acid molecule, wherein the structural nucleotide encodes any polypeptide or protein set forth in Table 1; which is operably linked to (C) a 3' non-translated sequence that functions in said cell to cause termination of transcription.

The present invention also provides a transformed cell having an exogenous nucleic acid molecule which comprises: (A) a promoter region which functions in said cell to cause the production of an mRNA molecule wherein the promoter region is selected from the group consisting of promoter sequences located within SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof; which is operably linked to (B) a structural nucleotide sequence encoding a polypeptide; which is operably linked to (C) a 3' non-translated sequence that functions in said cell to cause termination of transcription.

The present invention also provides a plant cell, a mammalian cell, a bacterial cell, an algal cell, an insect cell and a fungal cell transformed with an isolated nucleic acid molecule of the present invention.

The invention also provides isolated nucleic acid molecules comprising nucleotide sequences encoding polypeptides or proteins exhibiting insect inhibitory activity, wherein said activity is manifested by inhibiting the growth or development of, or contributing substantially to, or causing the death of a Coleopteran, a Dipteran, a Lepidopteran, a Hemipteran, a Hymenopteran, or a sucking and piercing insect or insect larvae thereof. Also provided are nucleotide sequences encoding novel proteins comprising polypeptides which augment the activity of polypeptides exhibiting insect inhibitory activity when fed to Coleopteran, Dipteran, Lepidopteran, Hemipteran, Hymenopteran, or sucking and piercing insects or insect larvae thereof.

The present invention also provides a method for using insect inhibitory proteins for controlling target insect pests, i.e. also known as insect pest control.

The present invention also provides a computer readable medium having recorded thereon one or more nucleotide sequences, wherein each of the nucleotide sequences is selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof.

The present invention also provides a computer readable medium having recorded thereon one or more of the nucleotide sequences encoding a protein or fragment thereof set forth in Table 1.

The present invention also provides a method for using the computer media of the present invention in isolating/identifying nucleic acids encoding insect inhibitory proteins, or proteins involved in biosynthesis of antibiotics.

The present invention also provides a method for identifying one or more genes encoding insect inhibitory proteins in the sequences of one or more plasmids of a *Bacillus thuringiensis*, wherein the method of the present invention comprises the steps of: a) isolating and purifying plasmid DNA; b) constructing a DNA library from the isolated and purified plasmid DNA; c) sequencing the DNA library to obtain a set of plasmid DNA sequences; d) comparing the set of DNA sequences with a set of chromosomal DNA sequences, wherein the set of chromosomal DNA sequences comprises the group consisting of SEQ ID No: 1 through SEQ ID No: 8283; e) identify common sequences, which are identified both in the set of plasmid DNA sequences and in the set of chromosomal DNA sequences; f) subtracting the common sequences from the set of plasmid DNA sequences to obtain a subtracted set of plasmid DNA sequences; g) assembling the subtracted set of DNA sequences to obtain contigs and sequence assemblies; h) determining open reading frames in the contigs and sequence assemblies; and h) identifying one or more genes encoding insect inhibitory proteins in the sequences of one or more plasmids of said *Bacillus thuringiensis*.

The present invention also provides a method for identifying plasmid DNA sequences of a *Bacillus* species, the method comprising the steps of a) identifying a *Bacillus* species strain which does

not contain plasmid DNA; b) generating a library of chromosomal genomic DNA from said *Bacillus* species strain which does not contain plasmid DNA; c) obtaining the nucleotide sequence of said chromosomal genomic DNA; d) identifying a *Bacillus* species strain which contains plasmid DNA; e) generating a library of said *Bacillus* species plasmid DNA; f) obtaining the nucleotide sequence of said plasmid DNA; g) subtracting any common sequences identified in the plasmid DNA which are also identified in the chromosomal genomic DNA; and h) constructing a contig sequence or sequence assemblies of said plasmid DNA, wherein said contig sequence or sequence assemblies comprise the plasmid DNA sequence of said *Bacillus* species.

## DETAILED DESCRIPTION OF THE INVENTION

### Agents of the Present Invention

#### Nucleic Acid Molecules

One aspect of the present invention relates to an isolated nucleic acid molecule having a first nucleotide sequence, wherein: (1) the first nucleotide sequence hybridizes under stringent conditions to a second nucleotide sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof, wherein the hybridizing portion of the second nucleotide sequence is at least 50 nucleotides in length; (2) the first nucleotide sequence is a portion of any sequence selected from the group consisting of SEQ ID NO:1 through SEQ ID NO: 8283; or (3) the first nucleotide sequence is the complement of (1) or (2).

The term "nucleic acid" means a single or double-stranded polymer of deoxyribonucleotide or ribonucleotide bases read from the 5' to the 3' end. Nucleic acids may also optionally contain synthetic, non-natural or altered nucleotide bases that permit correct read through by a polymerase and do not alter expression of a polypeptide encoded by that nucleic acid.

The term "an isolated nucleic acid" refers to a nucleic acid that is no longer accompanied by some of materials with which it is associated in its natural state or to a nucleic acid the structure of which is not identical to that of any of naturally occurring nucleic acid. Examples of an isolated nucleic acid include but are not limited to: (1) DNAs which have the sequence of part of a naturally occurring genomic DNA molecules but are not flanked by two coding sequences that flank that part of the molecule in the genome of the organism in which it naturally occurs; (2) a nucleic acid incorporated into a vector or into the genomic DNA of a prokaryote or eukaryote in a manner such that the resulting molecule is not identical to any naturally occurring vector or genomic DNA; (3) a separate molecule such as a cDNA, a genomic fragment, a fragment produced by polymerase chain reaction (PCR), or a restriction fragment;



(4) recombinant DNAs; and (5) synthetic DNAs. An isolated nucleic acid may also be comprised of one or more segments of cDNA, genomic DNA or synthetic DNA.

It is also contemplated by the inventors that the isolated nucleic acids of the present invention may also include known types of modifications, for example, labels which are known in the art, methylation, "caps", substitution of one or more of the naturally occurring nucleotides with an analog. Other known modifications include inter-nucleotide modifications, for example, those with uncharged linkages (methyl phosphonates, phosphotriesters, phosphoamidates, carbamates, etc.) and with charged linkages (phosphorothioates, phosphorodithioates, etc.), those containing pendant moieties, such as, proteins (including nucleases, toxins, antibodies, signal peptides, poly-L-lysine, etc.), those with intercalators (acridine, psoralen, etc.), those containing chelators (metals, radioactive metals, boron, oxidative metals, etc.), those containing alkylators, and those with modified linkages.

The term "nucleotide sequence" refers to both the sense and antisense strands of a nucleic acid as either individual single strands or in the duplex. It includes, but is not limited to, self-replicating plasmids, chromosomal sequences, and infectious polymers of DNA or RNA.

A nucleotide sequence is said to be the "complement" of another nucleotide sequence if they exhibit complete complementarity. As used herein, molecules are said to exhibit "complete complementarity" when every nucleotide of one of the sequences is complementary to a nucleotide of the other.

A "coding sequence" or "structural nucleotide sequence" is a nucleotide sequence which is translated into a polypeptide, usually via mRNA, when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a translation start codon at the 5'-terminus and a translation stop codon at the 3'-terminus. A coding sequence can include, but is not limited to, genomic DNA, cDNA, and recombinant polynucleotide sequences.

The term "recombinant DNAs" refers to DNAs that contains a genetically engineered modification through manipulation via mutagenesis, restriction enzymes, and the like.

The term "synthetic DNAs" refers to DNAs assembled from oligonucleotide building blocks that are chemically synthesized using procedures known to those skilled in the art. These building blocks are ligated and annealed to form DNA segments which are then enzymatically assembled to construct the entire DNA. "Chemically synthesized", as related to a sequence of DNA, means that the component nucleotides were assembled in vitro. Manual chemical synthesis of DNA may be accomplished using well established procedures, or automated chemical synthesis can be performed using one of a number of commercially available machines.

The term “stringent conditions” or “stringent hybridization conditions” refers to conditions under which a probe will hybridize to its target sequence, to a detectable greater degree than other sequences (e.g., at least 2 fold over background). Stringent conditions are sequence dependent and will be different in different circumstances. By controlling the stringency of the hybridization and/or washing conditions, target sequences can be identified which are 100% complementary to the probe (homologous probing). Alternatively, stringency conditions can be adjusted to allow some mismatching in sequences so that lower degrees of complementary to the probe are detected (heterologous probing).

Typically, stringent conditions will be those in which the salt concentration is less than about 1.5 M monovalent cation (e.g.,  $\text{Na}^+$ ), typically about 0.01 to 1.0 M monovalent cation concentration at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (e.g., 10 to 50 nucleotides) and at least about 60°C for long probes (e.g., greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide. Exemplary low stringency conditions include hybridization with a buffer solution of 30 to 35% formamide, 1 M NaCl, 1% SDS (sodium dodecyl sulfate) at 37°C, and a wash in 1X to 2X SSC (20X SSC = 3.0 M NaCl/0.3 M trisodium citrate) at 50 to 55°C. Exemplary moderate stringency conditions include hybridization in 40 to 45% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.5X to 1 X SSC at 55 to 60°C. Exemplary high stringency conditions include hybridization in 50% formamide, 1 M NaCl, 1% SDS at 37°C, and a wash in 0.1 X SSC at 60 to 65°C.

Specificity is typically the function of post-hybridization washes, the critical factors being the ionic strength and temperature of the final wash solution. For DNA-DNA hybrids, the  $T_m$  can be approximated from the equation of Meinkoth and Wahl, *Anal. Biochem.*, 138:267-284 (1984):  $T_m = 81.5^\circ\text{C} + 16.6 (\log M) + 0.41 (\%GC) - 0.61 (\% \text{ form}) - 500/L$ ; where M is the molarity of monovalent cations, %GC is the percentage of guanosine and cytosine nucleotides in the DNA, % form is the percentage of formamide in the hybridization solution, and L is the length of the hybrid in base pairs. The  $T_m$  is the temperature (under defined ionic strength and pH) at which 50% of a complementary target sequence hybridizes to a perfectly matched probe.  $T_m$  is reduced by about 1°C for each 1 % of mismatching; thus,  $T_m$  hybridization and/or wash conditions can be adjusted to hybridize to sequences of the desired identity. For example, if sequences with >90% identity are sought, the  $T_m$  can be decreased 10°C. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point ( $T_m$ ) for the specific sequence and its complement at a defined ionic strength and pH. However, severely stringent conditions can utilize a hybridization and/or wash at 1, 2, 3, or 4°C lower than the thermal melting point ( $T_m$ ); moderately stringent conditions can utilize a hybridization and/or wash at 6, 7, 8, 9, or 10°C lower than the thermal melting point ( $T_m$ ); low stringency conditions can utilize a hybridization

and/or wash at 11, 12, 13, 14, 15, or 20°C lower than the thermal melting point ( $T_m$ ). Using the equation, hybridization and wash compositions, and desired  $T_m$  those of ordinary skill will understand that variations in the stringency of hybridization and/or wash solutions are inherently described. If the desired degree of mismatching results in a  $T_m$  of less than 45°C (aqueous solution) or 32°C (formamide solution) it is preferred to increase the SSC concentration so that a higher temperature can be used. Appropriate stringent conditions are known to those skilled in the art or can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. (1989), 6.3.1-6.3.6.

The hybridization portion of the two hybridizing nucleic acids is usually at least 40 nucleotides in length, more usually at least about 75 nucleotides in length, more particularly at least 100 nucleotides in lengths. The hybridizing portion of the hybridizing nucleic acid is at least 80%, at least 90%, or at least 98% identical to the sequence of a portion of a sequence set forth in SEQ ID NO: 1 to SEQ ID NO: 8283.

Another aspect of the present invention relates to an isolated nucleic acid molecule comprising one or more open reading frames listed in Table 1. An "open reading frame" (ORF) is a region of a nucleotide sequence which encodes a polypeptide. This region may represent a portion of a coding sequence or a total coding sequence. Table 1 sets forth a list of open reading frames identified in the isolated nucleic acid molecules, wherein the open reading frames encode *Bacillus thuringiensis* proteins or polypeptide or fragments thereof which are homologues of known proteins or unknown proteins, or of tRNA's or rRNA's or fragments thereof which are homologues of known tRNA's or rRNA's.

Open reading frames in genomic sequences can be screened for the presence of protein homologues utilizing a number of different search algorithms that have been developed, one example of which is the suite of programs referred to as BLAST programs. There are five implementations of BLAST, three designed for nucleotide sequences queries (BLASTN, BLASTX, and TBLASTX) and two designed for protein sequence queries (BLASTP and TBLASTN) (Coulson, *Trends in Biotechnology* 12:76-80 (1994); Birren *et al.*, *Genome Analysis* 1:543-559 (1997)). Other examples of suitable programs that can be utilized are well known in the art. In addition, unidentified reading frames may be screened for by gene prediction software such as GenScan, which is located at <http://gnomic.stanford.edu/GENSCANW.html>. Novel genes, i.e., with no known homologs, can be predicted with the program GeneMark, which calculates the probability of a gene based on the presence of a gene-like 'grammar' in the DNA sequence (i.e., start and stop signals, and a significant open reading frame) and statistical analyses of protein-coding potential through biases in putative codon usage (see <http://genemark.biology.gatech.edu/GeneMark> for details).

The present invention also provides an isolated nucleic acid molecule comprising a first nucleotide sequence, wherein: (1) the first nucleotide sequence hybridizes under stringent conditions to a

second nucleotide sequence, wherein the hybridizing portion of the second nucleotide sequence encodes any polypeptide or protein set forth in Table 1; or (2) the first nucleotide sequence is the complement of (1) or (2).

In one embodiment, an isolated nucleic acid molecule comprises a nucleotide sequence, wherein the nucleotide sequence encodes any polypeptide or protein set forth in Table 1.

The term "polypeptide" or "protein" refers to a linear polymer composed of amino acids connected by peptide bonds.

By "substantial identical" or "substantially identical" as used in reference to two amino acid sequences, it is meant that one amino acid sequence is identical to the other amino acid sequence or has at least 50% sequence identity, at least 70% sequence identity, preferably at least 80%, more preferably at least 90%, and most preferably at least 95% identity when compared to the other amino acid sequence as a reference sequence using the programs described herein; preferably BLAST using standard parameters, as described below.

"Percentage of sequence identity" is determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity.

Polypeptides which are "substantially similar" share sequences as noted above except that residue positions which are not identical may differ by conservative amino acid changes. Conservative amino acid substitutions refer to the interchangeability of residues having similar side chains. "Conservative amino acid substitutions" refer to substitutions of one or more amino acids in a native amino acid sequence with another amino acid(s) having similar side chains, resulting in a silent change. Conserved substitutes for an amino acid within a native amino acid sequence can be selected from other members of the group to which the naturally occurring amino acid belongs. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and

methionine. Preferred conservative amino acids substitution groups are: valine-leucine, valine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, aspartic acid-glutamic acid, and asparagine-glutamine.

Optimal alignment of sequences for comparison can use any means to analyze sequence identity (homology) known in the art, e.g., by the progressive alignment method of termed "PILEUP" (Morrison, Mol. Biol. Evol. 14:428-441 (1997), as an example of the use of PILEUP); by the local homology algorithm of Smith & Waterman (Adv. Appl. Math. 2: 482 (1981)); by the homology alignment algorithm of Needleman & Wunsch (J. Mol. Biol. 48:443 (1970)); by the search for similarity method of Pearson (Proc. Natl. Acad. Sci. USA 85: 2444 (1988)); by computerized implementations of these algorithms (e.g., GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI); ClustalW (CLUSTAL in the PC/Gene program by Intelligenetics, Mountain View, California, described by, e.g., Higgins, Gene 73: 237-244 (1988); Corpet, Nucleic Acids Res. 16:10881-10890 (1988); Huang, Computer Applications in the Biosciences 8:155-165 (1992); and Pearson, Methods in Mol. Biol. 24:307-331 (1994); Pfam (Sonnhammer, Nucleic Acids Res. 26:322-325 (1998); TreeAlign (Hein, Methods Mol. Biol. 25:349-364 (1994); MES-ALIGN, and SAM sequence alignment computer programs; or, by manual visual inspection.

Another example of algorithm that is suitable for determining sequence similarity is the BLAST algorithm, which is described in Altschul et al, J. Mol. Biol. 215: 403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (NCBI), <http://www.ncbi.nlm.nih.gov/>; see also Zhang, Genome Res. 7:649-656 (1997) for the "PowerBLAST" variation. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence that either match or satisfy some positive valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul et al, J. Mol. Biol. 215: 403-410 (1990)). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T and X determine the sensitivity and speed of the alignment. The BLAST program uses as defaults a wordlength (W) of 11, the BLOSUM62 scoring matrix (see Henikoff, Proc. Natl. Acad. Sci. USA 89:10915-10919(1992)) alignments (B) of 50, expectation (E) of 10, M=5, N=-4, and a comparison of both strands.

The term BLAST refers to the BLAST algorithm which performs a statistical analysis of the similarity between two sequences; see, e.g., Karlin, Proc. Natl. Acad. Sci. USA 90:5873-5787 (1993). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001.

One skilled in the art will recognize that these values of sequence identity can be appropriately adjusted to determine corresponding sequence identity of two nucleotide sequences encoding the proteins of the present invention by taking into account codon degeneracy, conservative amino acid substitutions, reading frame positioning and the like. Substantial identity of nucleotide sequences for these purposes normally means sequence identity of at least 40%, preferably at least 60%, more preferably at least 90%, and most preferably at least 95%.

The term "codon degeneracy" refers to divergence in the genetic code permitting variation of the nucleotide sequence without effecting the amino acid sequence of an encoded polypeptide. The skilled artisan is well aware of the "codon-bias" exhibited by a specific host cell in usage of nucleotide codons to specify a given amino acid. Therefore, when synthesizing a gene for expression in a host cell, it is desirable to design the gene such that its frequency of codon usage approaches the frequency of preferred codon usage of the host cell.

The present invention also includes an isolated nucleic acid comprising a nucleotide sequence encoding a polypeptide set forth in Table 1, wherein the an amino acid sequence of the protein or polypeptide contains one or more conservative amino acid substitutions.

In a preferred embodiment of the present invention, the isolated nucleic acid molecule comprising a nucleotide sequence which encodes an insect inhibitory protein or polypeptide or fragment thereof.

In a preferred embodiment of the present invention, the isolated nucleic acid molecule comprising a nucleotide sequence encoding all or substantial portion of a sigma factor homologue listed in Table 2.

The term "insect inhibitory protein" refers to any polypeptide or protein or a substantial portion thereof that exhibits insect inhibitory activity, wherein said activity is manifested by inhibiting the growth or development of, or contributing substantially to, or causing the death of a Coleopteran, a Dipteran, a Lepidopteran, a Hemipteran, a Hymenopteran, or a sucking and piercing insect or insect larvae thereof.

The term "insect inhibitory protein" also refers to any polypeptide or protein with modified amino acid sequence, such as sequence which has been mutated, truncated, increased and the like and which maintains at least the insect inhibitory activity associated with the native protein. Accordingly, the isolated nucleic acids encoding those polypeptide or protein with such modification are also within the scope of the present invention.

In a preferred embodiment of the present invention, the isolated nucleic acid molecule comprising a nucleotide sequence which encodes whole or a portion of a protein homologue capable of conferring antibiotic resistance or resistance to heavy metals or other chemicals, wherein the nucleotide sequence is any sequence set forth in SEQ ID NOs: 33, 98, 145, 162, 180, 204, 275, 298, 361, 397, 421, 423, 579, 613, 624, 692, 726, 862, 930, 950, 986, 995, 1005, 1023, 1130, 1188, 1190, 1208, 1226, 1227, 1240, 1246, 1246, 1257, 1272, 1302, 1339, 1355, 1374, 1393, 1426, 1460, 1471, 1526, 1854, 1914, 1923, 2151, 2179, 2211, and 2304 (Table 4).

In a preferred embodiment of the present invention, the isolated nucleic acid molecule comprising a nucleotide sequence which encodes whole or a portion of a transposon or transposase homologue, wherein the nucleotide sequence is any sequence set forth in SEQ ID Nos: 2, 64, 226, 379, 383, 387, 410, 416, 546, 555, 603, 642, 644, 660, 691, 691, 781, 799, 980, 1002, 1045, 1072, 1098, 1190, 1207, 1214, 1252, 1273, 1275, 1305, 1317, 1330, 1340, 1353, 1354, 1362, 1378, 1378, 1380, 1383, 1386, 1386, 1388, 1391, 1392, 1549, 1573, 1611, 1698, 1725, 1739, 1804, 1869, 1902, 1965, 2041, 2049, 2130, 2135, 2153, 102, 1340, 1795, 1797, 1989, 2055, 2057, 2248, 14, 296, 722, 834, 834, and 999 (Table 3).

In a preferred embodiment of the present invention, the isolated nucleic acid molecule comprising a nucleotide sequence encodes whole or a portion of a toxin or toxin homologue listed in Table 5.

In a preferred embodiment of the present invention, the isolated nucleic acid molecule encodes a *B. thuringiensis* protein or fragment thereof that is a homologue of another *Bacillus* protein. In another preferred embodiment of the present invention, the nucleic acid molecule encodes a *B. thuringiensis* protein or fragment thereof that is a homologue of a fungal protein. In another preferred embodiment of the present invention, the nucleic acid molecule encodes a *B. thuringiensis* protein or fragment thereof that is a homologue of a plant protein. In another preferred embodiment of the present invention, the nucleic acid molecule encodes a *B. thuringiensis* protein or fragment thereof that is a homologue of mammalian protein. In another preferred embodiment of the present invention, a *B. thuringiensis* protein or fragment thereof of the present invention is a homologue of a non-*Bacillus* bacterial protein. In another preferred embodiment of the present invention, the nucleic acid molecule encodes a *B. thuringiensis* protein or fragment thereof that is a homologue of an algal protein.

In a preferred embodiment of the present invention, the nucleic acid molecule of the present invention encodes a *B. thuringiensis* homologue protein or fragment thereof where the *B. thuringiensis* homologue protein exhibits a BLASTP probability score of greater than 1E-12, preferably a BLASTP probability score of between about 1E-30 and about 1E-12, even more preferably a BLASTP probability score of greater than 1E-30 with its homologue. In a preferred embodiment of the present invention, the nucleic acid molecule of the present invention encodes a *B. thuringiensis* homologue protein or fragment thereof exhibits an *aat\_nap* score of less than 200, preferably an *aat\_nap* score of between about 200 to about 400, even more preferably an *aat\_nap* score of greater than 400.

It is understood that certain amino acids may be substituted for other amino acids in a protein structure without appreciable loss of interactive binding capacity with structures such as, for example, antigen-binding regions of antibodies or binding sites on substrate molecules. Because it is the interactive capacity and nature of a protein that defines that protein's biological functional activity, certain amino acid sequence substitutions can be made in a protein sequence and, of course, its underlying DNA coding sequence and, nevertheless, obtain a protein with like properties. It is thus contemplated by the inventors that various changes may be made in the peptide sequences of the proteins or fragments of the present invention, or corresponding DNA sequences that encode said peptides, without appreciable loss of their biological utility or activity. It is understood that codons capable of coding for such amino acid changes are known in the art.

In making such changes, the hydropathic index of amino acids may be considered. The importance of the hydropathic amino acid index in conferring interactive biological function on a protein is generally understood in the art (Kyte and Doolittle, *J. Mol. Biol.* 157, 105-132 (1982)). It is accepted that the relative hydropathic character of the amino acid contributes to the secondary structure of the resultant protein, which in turn defines the interaction of the protein with other molecules, for example, enzymes, substrates, receptors, DNA, antibodies, antigens, and the like.

Each amino acid has been assigned a hydropathic index on the basis of its hydrophobicity and charge characteristics (Kyte and Doolittle, 1982); these are isoleucine (+4.5), valine (+4.2), leucine (+3.8), phenylalanine (+2.8), cysteine/cystine (+2.5), methionine (+1.9), alanine (+1.8), glycine (-0.4), threonine (-0.7), serine (-0.8), tryptophan (-0.9), tyrosine (-1.3), proline (-1.6), histidine (-3.2), glutamate (-3.5), glutamine (-3.5), aspartate (-3.5), asparagine (-3.5), lysine (-3.9), and arginine (-4.5).

In making such changes, the substitution of amino acids whose hydropathic indices are within  $\pm 2$  is preferred, those which are within  $\pm 1$  are particularly preferred, and those within  $\pm 0.5$  are even more particularly preferred.



It is also understood in the art that the substitution of like amino acids can be made effectively on the basis of hydrophilicity. U.S. Patent 4,554,101, incorporated herein by reference in its entirety, states that the greatest local average hydrophilicity of a protein, as govern by the hydrophilicity of its adjacent amino acids, correlates with a biological property of the protein.

As detailed in U.S. Patent 4,554,101, the following hydrophilicity values have been assigned to amino acid residues: arginine (+3.0), lysine (+3.0), aspartate (+3.0 $\pm$ 1), glutamate (+3.0 $\pm$ 1), serine (+0.3), asparagine (+0.2), glutamine (+0.2), glycine (0), threonine (-0.4), proline (-0.5 $\pm$ 1), alanine (-0.5), histidine (-0.5), cysteine (-1.0), methionine (-1.3), valine (-1.5), leucine (-1.8), isoleucine (-1.8), tyrosine (-2.3), phenylalanine (-2.5), and tryptophan (-3.4).

In making such changes, the substitution of amino acids whose hydrophilicity values are within  $\pm 2$  is preferred, those which are within  $\pm 1$  are particularly preferred, and those within  $\pm 0.5$  are even more particularly preferred.

Another aspect of the present invention relates to a class of isolated nucleic acid molecules comprising promoter sequences or regulatory elements, particularly those found within SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof.

The term "promoter sequence" means a nucleotide sequence that is capable of, when located in *cis* to a structural nucleotide sequence encoding a polypeptide or protein, functioning in a way that directs expression of one or more mRNA molecules that encodes the polypeptide or protein. Such promoter regions are typically found upstream of the trinucleotide ATG sequence at the start site of a protein coding region. Promoter sequences can also include sequences from which transcription of transfer RNA (tRNA) or ribosomal RNA (rRNA) sequences are initiated. Transcription involves the synthesis of an RNA chain representing one strand of a DNA duplex. By "representing" it is meant that the RNA is identical in sequence with one strand of the DNA; it is complementary to the other DNA strand, which provides the template for its synthesis. Transcription takes place by the usual process of complementary base pairing, catalyzed and scrutinized by the enzyme RNA polymerase. The reaction can be divided into three stages described as initiation, elongation and termination. Initiation begins with the binding of RNA polymerase to the double stranded (DS or ds) DNA. The sequence of DNA required for the initiation reaction defines the promoter. The site at which the first nucleotide is incorporated is called the startsite or startpoint of transcription. Elongation describes the phase during which the enzyme moves along the DNA and extends the growing RNA chain. Elongation involves the disruption of the DNA double stranded structure in which a transiently unwound region exists as a hybrid RNA-DNA duplex and a displaced single strand of DNA. Termination involves recognition of the point at which no further bases should be added to the chain. To terminate transcription, the formation of phosphodiester

bonds must cease and the transcription complex must come apart. When the last base is added to the RNA chain, the RNA-DNA hybrid is disrupted, the DNA reforms into a duplex state, and the RNA polymerase enzyme and RNA molecule are both released from the DNA. The sequence of DNA required for the termination reaction is called the terminator.

5 Generally, for bacteria the optimal promoter is a sequence consisting of a -35 hexamer separated by about 17 base pairs from a -10 hexamer and lies from about 7 to about 10 base pairs upstream of the startpoint of transcription, but these sequences can vary among and between sequences which are recognized by the RNA polymerase. The startpoint of transcription generally lies from about 20 to about 50 base pairs upstream of the startpoint of translation of one or more open reading frames which  
10 comprise the entire length of an mRNA transcript. Some promoters can be recognized by RNA polymerase alone and in these cases, an accessible promoter will always be transcribed. Promoter availability may be determined by extraneous proteins, which either may act directly at the promoter to block access by RNA polymerase, or may function indirectly by controlling the structure of the genome in the region. Other promoters are not by themselves adequate to support transcription initiation and thus  
15 ancillary protein and or RNA factors are required to further initiation. The additional protein or RNA factors usually act by recognizing sequences of DNA that are close to, or overlap with, the sequence bound by RNA polymerase itself. Additionally, some of these ancillary factors must touch and concern the RNA polymerase in order to effect efficient transcription initiation as well as transcription elongation.

Promoters in particular in *Bacillus* species are highly regulated by the appearance and  
20 disappearance of accessory factors known as sigma factors which touch and concern both the DNA sequences flanking the promoter site as well as the RNA polymerase in order to effect efficient transcription initiation and elongation. Such factors are required for effecting transcription from various classes of promoters and along with other factors expressed from *spoO* genes affect both the temporal regulation of expression from promoters as well as the spatial distribution of patterns of expression within  
25 the cell during differentiation and development of the *Bacillus* cell from spore activation, vegetative growth and proliferation, and sporulation. Examples of sigma factors which are known to function in this manner are  $\sigma^A$ ,  $\sigma^B$ ,  $\sigma^C$ ,  $\sigma^D$ ,  $\sigma^E$ ,  $\sigma^F$ ,  $\sigma^G$ ,  $\sigma^H$ ,  $\sigma^I$ , and  $\sigma^K$ . In addition, the sporulation factors which function for temporal and spatial regulatory gene expression include gene products in the classes of SpoO, SpoOI, and SpoOII.

30 Thus, in *Bacillus*, in many circumstances a consensus promoter may be sufficient for expression, however additional sequences may act to further regulate expression, for example, in response to biochemical, developmental or environmental signals. In a preferred embodiment of the present

invention, the promoter is upstream of a nucleic acid sequence that encodes a *B. thuringiensis* protein or fragment thereof.

Promoters of the present invention can be included within sequences up to 10 kb upstream of the trinucleotide ATG sequence at the start site of a protein coding region, tRNA, or rRNA. Promoters of the present invention can preferably be included within sequences up to 5 kb upstream of the trinucleotide ATG sequence at the start site of a protein coding region, tRNA or rRNA. Promoters of the present invention can more preferably be included within sequences up to 2kb upstream of the trinucleotide ATG sequence at the start site of a protein coding region, tRNA or rRNA. Promoters of the present invention can most preferably be included within sequences up to 500 bp upstream of the trinucleotide ATG sequence at the start site of a protein, tRNA, or rRNA coding region. While in many circumstances a 300 bp promoter may be sufficient for expression, additional sequences may act to further regulate expression, for example, in response to biochemical, developmental or environmental signals. In a preferred embodiment of the present invention, the promoter is upstream of an nucleic acid sequence that encodes a *Bacillus thuringiensis* protein or fragment thereof.

The term "regulatory element" is intended to mean a series of nucleotides that determines if, when, and at what level a particular gene is expressed. Regulatory DNA sequences specifically interact with regulatory or other proteins. Many regulatory elements act in *cis* ("*cis* elements") and are believed to affect DNA topology, producing local conformations that selectively allow or restrict access of RNA polymerase to the DNA template or that facilitate selective opening of the double helix at the site of transcriptional initiation, i.e., the transcriptional startsite referred to above. *Cis* elements occur within, near to, adjacent to, or at a distance from a particular promoter, but remain linked to the promoter sequence along the sequence of phosphodiester bonds which comprise the nucleotide sequence within which the promoter resides. *Cis* elements are not limited to promoters, but may be imparted to RNA sequences derived from transcription from DNA sequences of the present invention, wherein such RNA *cis* elements are involved in post transcriptional regulation of gene expression. For example, elements which are known as inverted repeat sequences can assist in the formation of hairpin structures which prevent, inhibit, or otherwise modulate the translational efficiency of the RNA sequence, or which regulate the survival of the RNA sequence. Other elements may function to bind ribosomes or components which enhance or suppress translational efficiency. *Cis* elements can be identified using known *cis* elements as a target sequence or target motif in the BLAST. Promoters of the present invention include homologues of *cis* elements known to effect gene regulation that show homology with the nucleic acid molecules of the present invention.

The isolated nucleic acid molecules of the present invention also include nucleic acid sequences that encode ribosomal RNA (rRNA), transfer RNA (tRNA) molecules, or other nucleic acid molecules which function to regulate gene expression, transcription, translation by acting alone or in combination with other cellular components in activating, inhibiting, terminating or anti-terminating gene expression functions, or by acting alone or in combination with other structural molecules to form or assist in the formation of said structural molecules.

It is contemplated by the inventors that the isolated nucleic acid molecules of the present invention also include those comprising a substantial portion of a nucleotide sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof.

A "substantial portion" of a nucleotide sequence comprises enough of the sequence to afford specific identification and/or isolation of a nucleic acid fragment comprising the sequence. In general, gene specific oligonucleotide probes comprising 20-30 contiguous nucleotides may be used in sequence-dependent methods of gene identification (e.g., Southern hybridization) and isolation (e.g., in situ hybridization of bacterial colonies or bacteriophage plaques). In addition, short oligonucleotides of 12-15 bases may be used as amplification primers in PCR in order to obtain a particular nucleic acid fragment comprising the primers. The skilled artisan, having the benefit of the sequences as reported herein, may now use all or a substantial portion of the disclosed sequences for purposes known to those skilled in this art. Accordingly, the instant invention comprises the complete sequences as reported in the accompanying Sequence Listing, as well as substantial portions of those sequences as defined above.

It is also contemplated by the inventors that the isolated nucleic acid molecules of the present invention also include known types of modifications, for example, labels which are known in the art, methylation, "caps", substitution of one or more of the naturally occurring nucleotides with an analog. Other known modifications include internucleotide modifications, for example, those with uncharged linkages (methyl phosphonates, phosphotriesters, phosphoamidates, carbamates, etc.) and with charged linkages (phosphorothioates, phosphorodithioates, etc.), those containing pendant moieties, such as, proteins (including nucleases, metabolic toxins, antibodies, signal peptides, poly-L-lysine, etc.), those with intercalators (acridine, psoralen, etc.), those containing chelators (metals, radioactive metals, boron, oxidative metals, etc.), those containing alkylators, and those with modified linkages.

The nucleic acids of the present invention may be used to isolate nucleic acids encoding homologous proteins from the same or other species. Isolation of homologous genes using sequence-dependent protocols is well known in the art. Examples of sequence-dependent protocols include, but are not limited to, methods of nucleic acid hybridization, and methods of DNA and RNA amplification as

exemplified by various uses of nucleic acid amplification technologies (e.g., polymerase chain reaction, ligase chain reaction).

For example, genes encoding homologous proteins, either as cDNAs or genomic DNAs, could be isolated directly by using all or a portion of the nucleic acids of the present invention as DNA hybridization probes to screen cDNA or genomic libraries from any desired organism employing methodology well known to those skilled in the art. Methods for forming such libraries are well known in the art. Specific oligonucleotide probes based upon the nucleic acids of the present invention can be designed and synthesized by methods known in the art. Moreover, the entire sequences of the nucleic acids can be used directly to synthesize DNA probes by methods known to the skilled artisan such as random primer DNA labeling, nick translation, or end-labeling techniques, or RNA probes using available in vitro transcription systems. In addition, specific primers can be designed and used to amplify a part or all of the sequences. The resulting amplification products can be labeled directly during amplification reactions or labeled after amplification reactions, and used as probes to isolate full length cDNA or genomic DNAs under conditions of appropriate stringency.

Alternatively, the nucleic acids of interest can be amplified from nucleic acid samples using amplification techniques. For instance, the disclosed nucleic acids may be used to define a pair of primers that can be used with the polymerase chain reaction (Mullis, *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* 51:263-273 (1986); Erlich *et al.*, EP 50,424; EP 84,796, EP 258,017, EP 237,362; Mullis, EP 201,184; Mullis *et al.*, US 4,683,202; Erlich, US 4,582,788; and Saiki, R. *et al.*, US 4,683,194) to amplify and obtain any desired nucleic acid or fragment directly from mRNA, from cDNA, from genomic libraries or cDNA libraries. PCR and other in vitro amplification methods may also be useful, for example, to clone nucleic acid sequences that code for proteins to be expressed, to make nucleic acids to use as probes for detecting the presence of the desired mRNA in samples, for nucleic acid sequencing, or for other purposes.

In addition, two short segments of the nucleic acids of the present invention may be used in polymerase chain reaction protocols, for example, the RACE protocol (Frohman *et al.*, *Proc. Natl. Acad. Sci. USA* 85:8998 (1988)), to amplify longer nucleic acids encoding homologous genes from DNA or RNA from other sources.

Nucleic acids of interest may also be synthesized, either completely or in part, especially where it is desirable to provide plant-preferred sequences, by well-known techniques as described in the technical literature. See, e.g., Carruthers *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* 47:411-418 (1982), and Adams *et al.*, *J. Am. Chem. Soc.* 105:661 (1983). Thus, all or a portion of the nucleic acids of the present invention may be synthesized using codons preferred by a selected plant host. Plant-preferred codons

may be determined, for example, from the codons used most frequently in the proteins expressed in a particular plant host species. Other modifications of the gene sequences may result in mutants having slightly altered activity.

Availability of the nucleotide sequences encoding *Bacillus thuringiensis* proteins facilitates immunological screening of DNA expression libraries. Synthetic polypeptides representing portions of the amino acid sequences of *Bacillus thuringiensis* proteins may be synthesized. These polypeptides can be used to immunize animals to produce polyclonal or monoclonal antibodies with specificity for polypeptides or proteins comprising the amino acid sequences. These antibodies can be then be used to screen expression libraries to isolate genes of interest (Lemer, Adv. ImmunoL 36: 1 (1984); Sambrook et al., Molecular Cloning: A Laboratory Manual; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, (1989)). It is understood that people skilled in the art are familiar with the standard resource materials which describe specific conditions and procedures for the construction, manipulation and isolation of antibodies (see, for example, Harlow and Lane, In *Antibodies: A Laboratory Manual*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1988)).

Another aspect of the present invention relates to a method for obtaining a nucleic acid comprising a nucleotide sequence encoding a *Bacillus thuringiensis* protein homologue listed in Table 1. In a preferred embodiment, the method of the present invention for obtaining a nucleic acid encoding all or a substantial portion of the amino acid sequence of a *Bacillus thuringiensis* protein homologue comprising: (a) probing an expression library with a hybridization probe comprising a nucleotide sequence encoding a polypeptide listed in Table 1; (b) identifying a DNA clone that hybridizes to the hybridization probe; (c) isolating the DNA clone identified in step (b); and (d) sequencing the DNA fragment that comprises the clone isolated in step (c) wherein the sequenced nucleic acid molecule encodes all or a substantial portion of the amino acid sequence of the *Bacillus thuringiensis* protein homologue.

In another preferred embodiment, the method of the present invention for obtaining a nucleic acid molecule encoding a substantial portion of an amino acid sequence of a *Bacillus thuringiensis* protein homologue comprising: (a) synthesizing a first and a second oligonucleotide primers corresponding to a portion of the coding sequence of a second nucleic acid molecule set forth in SEQ ID NO: 1 through SEQ ID NO: 8283; and (b) amplifying a DNA insert present in a cloning vector using the first and second oligonucleotide primers of step (a) wherein the amplified nucleic acid molecule encodes all or a substantial portion of the amino acid sequence of the *Bacillus thuringiensis* protein homologue.

### Protein and Polypeptide Molecules

The present invention, in another aspect, provides a substantially purified protein or polypeptide molecule comprising an amino acid sequence, wherein the amino acid sequence is defined as follows: (1) the amino acid sequence is encoded by a first nucleotide sequence which specifically hybridizes to the complement of a second nucleotide sequence set forth in SEQ ID NO: 1 through SEQ ID NO: 8283; or (2) the amino acid sequence is encoded by a third nucleotide sequence that is at least 50% identical to all or a substantial portion of a coding sequence located within SEQ ID NO: 1 through SEQ ID NO: 8283. In alternative embodiments, the third nucleotide sequence is at least 55% identical, at least 60% identical, at least 65% identical, at least 70% identical, at least 75% identical, at least 80% identical, at least 85% identical, at least 90% identical, at least 95% identical to all or a substantial portion of a coding sequence located within SEQ ID NO: 1 through SEQ ID NO: 8283. In a preferred embodiment, the third nucleotide sequence is 100% identical to all or a substantial portion of a coding sequence located within SEQ ID NO: 1 through SEQ ID NO: 8283.

The term "substantially purified protein or polypeptide molecule" refers to a protein or polypeptide molecule separated from substantially all other molecules normally associated with it in its native state. More preferably a substantially purified protein or polypeptide molecule is the predominant species present in a preparation. A substantially purified molecule may be greater than 60% free, preferably 75% free, more preferably 90% free, and most preferably 95% free from the other molecules (exclusive of solvent) present in the natural mixture.

It is well known in the art that proteins or polypeptides may undergo modification, including post-translational modifications, such as, but not limited to, disulfide bond formation, glycosylation, phosphorylation, or oligomerization. Thus, as used herein, the term "protein molecule" or "polypeptide molecule" includes any protein molecule that is modified by any biological or non-biological process. The terms "amino acid" and "amino acids" refer to all naturally occurring L-amino acids. This definition is meant to include norleucine, ornithine, homocysteine, and homoserine.

The polypeptides or proteins of the present invention may be produced via chemical synthesis, or more preferably, by expression in a suitable bacterial or eukaryotic host. Suitable methods for expression are described by Sambrook, *et al.*, (In: *Molecular Cloning, A Laboratory Manual, 2nd Edition*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1989)), or similar texts.

The polypeptides or protein molecule of the present invention may also include fusion protein or polypeptide molecules. A protein or polypeptide molecule that comprises one or more additional polypeptide regions not derived from that protein molecule is a "fusion" protein or polypeptide molecule. Such molecules may be derivatized to contain carbohydrate or other moieties (such as keyhole limpet

hemocyanin, etc.). Fusion protein or polypeptide molecules of the present invention are preferably produced via recombinant means.

The protein or polypeptide molecules of the present invention may also include protein or polypeptide molecules encoded by all or a substantial portion of protein-encoding sequences in SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof or, fragments or fusions thereof in which conservative, non-essential, or not relevant, amino acid residues have been added, replaced, or deleted. An example of such a homologue is the homologue protein from different strains or species. Such a homologue can be obtained by any of a variety of methods. For example, as indicated above, one or more of the disclosed sequences (all or a substantial portion of the protein-encoding sequences in SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof) will be used to define a pair of primers that may be used to isolate the homologue-encoding nucleic acid molecules from any desired species. Such molecules can be expressed to yield homologues by recombinant means.

#### **Antibodies**

Another aspect of the present invention concerns antibodies, single-chain antigen binding molecules, or other proteins that specifically bind to one or more of the protein or polypeptide molecules of the present invention and their homologues, fusions or fragments. Such antibodies may be used to quantitatively or qualitatively detect the protein or polypeptide molecules of the present invention. As used herein, an antibody or polypeptide is said to "specifically bind" to a protein or polypeptide molecule of the present invention if such binding is not competitively inhibited by the presence of non-related molecules. In a preferred embodiment the antibodies of the present invention bind to protein or polypeptide molecules of the present invention, in a more preferred embodiment of the antibodies of the present invention bind to protein or polypeptide molecules derived from *Bacillus thuringiensis*.

Nucleic acid molecules that encode all or part of the protein or polypeptide of the present invention can be expressed, via recombinant means, to yield protein or polypeptides that can in turn be used to elicit antibodies that are capable of binding the expressed protein or polypeptide. Such antibodies may be used in immunoassays for that protein or polypeptide. Such protein or polypeptide-encoding molecules, or their fragments may be "fusion" molecules (*i.e.*, a part of a larger nucleic acid molecule) such that, upon expression, a fusion protein is produced. It may be desirable to derivatize the obtained antibodies, for example with a ligand group (such as biotin) or a detectable marker group (such as a fluorescent group, a radioisotope or an enzyme). Such antibodies may be used in immunoassays for that protein. In a preferred embodiment, such antibodies can be used to screen DNA expression libraries to isolate clones containing full-length insert of genes (Lemer, Adv. ImmunoL 36: 1 (1984); Sambrook et



al., Molecular Cloning: A Laboratory Manual; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, (1989)).

The antibodies that specifically bind proteins and protein fragments of the present invention may be polyclonal or monoclonal, and may comprise intact immunoglobulins, or antigen binding portions of immunoglobulins (such as (F(ab'), F(ab')<sub>2</sub> fragments), or single-chain immunoglobulins producible, for example, via recombinant means). It is understood that practitioners are familiar with the standard resource materials which describe specific conditions and procedures for the construction, manipulation and isolation of antibodies (see, for example, Harlow and Lane, *Antibodies: A Laboratory Manual*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1988)).

In a preferred embodiment, the antibodies of the present invention specifically bind to one or more of the insect inhibitory polypeptides or proteins of the present invention. Such antibodies may be used to detect the presence of such insect inhibitory polypeptides or proteins in a sample.

The present invention also provide a method for detecting an insect inhibitory polypeptide or protein in a biological sample, the method generally comprising: (1) obtaining a biological sample; (2) contacting the sample with an antibody that specifically binds to the polypeptide or protein, under conditions effective to allow the formation of complexes; and (3) detecting the complexes so formed.

#### **Microbial Constructs and Transformed microbial Cells**

The nucleotide sequences of the present invention may be introduced into a wide variety of prokaryotic and eukaryotic microorganism hosts to express the *B. thuringiensis* polypeptide or protein of interest, particularly the insect inhibitory polypeptides or proteins of the present invention. The term "microorganism" includes prokaryotic and eukaryotic microbial species such as bacteria and fungi. Fungi include yeast and filamentous fungi. Illustrative prokaryotes, both Gram-negative and Gram-positive, include *Enterobacteriaceae*, such as *Escherichia*, *Erwinia*, *Shigella*, *Salmonella*, and *Proteus*; *Bacillaceae*; *Rhizobiceae*, such as *Rhizobium*; *Spirillaceae*, such as photobacterium, *Zymomonas*, *Serratia*, *Aeromonas*, *Vibrio*, *Desulfovibrio*, *Spirillum*; *Lactobacillaceae*; *Pseudomonadaceae*, such as *Pseudomonas* and *Acetobacter*; *Azotobacteraceae*, *Actinomycetales*, and *Nitrobacteraceae*. Among eukaryotes are fungi, such as *Phycomycetes* and *Ascomycetes*, which includes yeast, such as *Saccharomyces* and *Schizosaccharomyces*; and *Basidiomycetes* yeast, such as *Rhodotorula*, *Aureobasidium*, *Sporobolomyces*, and the like.

For the purpose of plant protection against insects, a large number of microorganisms known to inhabit the phylloplane (the surface of the plant leaves) and/or the rhizosphere (the soil surrounding plant

roots) of a wide variety of important crops may also be desirable host cells for manipulation, propagation, storage, delivery and/or mutagenesis of the disclosed recombinant constructs. These microorganisms include bacteria, algae, and fungi. Of particular interest are microorganisms, such as bacteria, *e.g.*, genera *Bacillus* (including the species and subspecies *B. thuringiensis kurstaki* HD-1, *B. thuringiensis kurstaki* HD-73, *B. thuringiensis sotto*, *B. thuringiensis berliner*, *B. thuringiensis thuringiensis*, *B. thuringiensis tolworthi*, *B. thuringiensis dendrolimus*, *B. thuringiensis alesti*, *B. thuringiensis galleriae*, *B. thuringiensis aizawai*, *B. thuringiensis subtoxicus*, *B. thuringiensis entomocidus*, *B. thuringiensis tenebrionis* and *B. thuringiensis san diego*); *Pseudomonas*, *Erwinia*, *Serratia*, *Klebsiella*, *Zanthomonas*, *Streptomyces*, *Rhizobium*, *Rhodopseudomonas*, *Methylophilus*, *Agrobacterium*, *Acetobacter*, *Lactobacillus*, *Arthrobacter*, *Azotobacter*, *Leuconostoc*, and *Alcaligenes*; fungi, particularly yeast, *e.g.*, genera *Saccharomyces*, *Cryptococcus*, *Kluyveromyces*, *Sporobolomyces*, *Rhodotorula*, and *Aureobasidium*. Of particular interest are such phytosphere bacterial species as *Pseudomonas syringae*, *Pseudomonas fluorescens*, *Serratia marcescens*, *Acetobacter xylinum*, *Agrobacterium tumefaciens*, *Rhodobacter sphaeroides*, *Xanthomonas campestris*, *Rhizobium melioli*, *Alcaligenes eutrophus*, and *Azotobacter vinlandii*; and phytosphere yeast species such as *Rhodotorula rubra*, *R. glutinis*, *R. marina*, *R. aurantiaca*, *Cryptococcus albidus*, *C. diffluens*, *C. laurentii*, *Saccharomyces rosei*, *S. pretoriensis*, *S. cerevisiae*, *Sporobolomyces roseus*, *S. odor*, *Kluyveromyces veronae*, and *Aureobasidium pollulans*.

It is well known that exogenous nucleic acids encoding polypeptides of interest can be introduced into a microbial host cell, such as a bacterial cell or a fungal cell, using a recombinant construct. The present invention also relates to a fungal or bacterial recombinant construct comprising a structural nucleotide sequence encoding a *B. thuringiensis* protein or polypeptide. In a preferred embodiment, the structural nucleotide sequence encodes an insect inhibitory protein or polypeptide of the present invention. The present invention also relates to a bacterial or fungal cell comprising a bacterial or fungal recombinant vector. The present invention also relates to methods for obtaining a recombinant bacterial or fungal host cell, comprising introducing into a bacterial or fungal host cell an exogenous nucleic acid molecule.

The bacterial recombinant vector may be any vector which can be conveniently subjected to recombinant DNA procedures. The choice of a vector will typically depend on the compatibility of the vector with the bacterial host cell into which the vector is to be introduced. The vector may be a linear or a closed circular plasmid. The bacterial recombinant vector may be a linear or a closed circular plasmid. The vector system may be a single vector or plasmid or two or more vectors or plasmids which together contain the total DNA to be introduced into the genome of the bacterial host. In addition, the bacterial vector may be an expression vector. Nucleic acid molecules encoding *B. thuringiensis* proteins or

polypeptide can, for example, be suitably inserted into a replicable vector for expression in a bacterium under the control of a suitable promoter for that bacterium. Many vectors are available for this purpose, and selection of the appropriate vector will depend mainly on the size of the nucleic acid to be inserted into the vector and the particular host cell to be transformed with the vector. Each vector contains various components depending on its function (amplification of DNA or expression of DNA) and the particular host cell with which it is compatible. The vector components for bacterial transformation generally include, but are not limited to, one or more of the following: a signal sequence, an origin of replication, one or more selectable marker genes, and an inducible promoter allowing the expression of exogenous DNA.

In general, plasmid vectors containing replicon and control sequences that are derived from species compatible with the host cell are used in connection with bacterial hosts. The vector ordinarily carries a replication site, as well as marking sequences that are capable of providing phenotypic selection in transformed cells. For example, *E. coli* is typically transformed using pBR322, a plasmid derived from an *E. coli* species (see, e.g., Bolivar *et al.*, *Gene* 2:95 (1977)). pBR322 contains genes for ampicillin and tetracycline resistance and thus provides easy means for identifying transformed cells. The pBR322 plasmid, or other microbial plasmid or phage, also generally contains, or is modified to contain, promoters that can be used by the microbial organism for expression of the selectable marker genes.

Nucleic acid molecules encoding *B. thuringiensis* proteins or polypeptides may be expressed not only directly, but also as a fusion with another polypeptide, preferably a signal sequence or other polypeptide having a specific cleavage site at the N-terminus of the mature polypeptide. In general, the signal sequence may be a component of the vector, or it may be a part of the polypeptide encoding DNA that is inserted into the vector. The heterologous signal sequence selected should be one that is recognized and processed (i.e., cleaved by a signal peptidase) by the host cell. For bacterial host cells that do not recognize and process the native polypeptide signal sequence, the signal sequence is substituted by a bacterial signal sequence selected, for example, from the group consisting of the alkaline phosphatase,  $\beta$ -lactamase, , or heat-stable enterotoxin II leaders and the like.

Both expression and cloning vectors contain a nucleotide sequence that enables the vector to replicate in one or more selected host cells. Generally, in cloning vectors this sequence is one that enables the vector to replicate independently of the host chromosomal DNA, and includes origins of replication or autonomously replicating sequences. Such sequences are well known for a variety of bacteria.

Expression and cloning vectors also generally contain a selection gene, also termed a selectable marker. This gene encodes a protein necessary for the survival or growth of transformed host cells grown

in a selective culture medium. Host cells not transformed with the vector containing the selection gene will not survive in the culture medium. Typical selection genes encode proteins that (a) confer resistance to antibiotics or other toxins, e.g., ampicillin, neomycin, methotrexate, or tetracycline, (b) complement auxotrophic deficiencies, or (c) supply critical nutrients not available from complex media, e.g., the gene encoding D-alanine racemase for *Bacilli*. One example of a selection scheme utilizes a drug to arrest growth of a host cell. Those cells that are successfully transformed with a heterologous protein homologue or fragment thereof produce a protein conferring drug resistance and thus survive the selection regimen.

The expression vector for producing a polypeptide can also contains an inducible promoter that is recognized by the host bacterial organism and is operably linked to the nucleic acid encoding, for example, the nucleic acid molecule encoding the *B. thuringiensis* protein or polypeptide of interest. Inducible promoters suitable for use with bacterial hosts include the  $\beta$ -lactamase, *E. coli*  $\lambda$  phage P<sub>L</sub> and P<sub>R</sub>, and *E. coli* galactose, arabinose, alkaline phosphatase, tryptophan (trp), and lactose operon promoter systems and variations thereof (Chang *et al.*, *Nature* 275:615 (1978); Goeddel *et al.*, *Nature* 281:544 (1979); Guzman *et al.*, *J. Bacteriol.* 174:7716-7728 (1992); Goeddel, *Nucleic Acids Res.* 8:4057 (1980); EP 36,776) and hybrid promoters such as the tac promoter (deBoer *et al.*, *Proc. Natl. Acad. Sci. (USA)* 80:21-25 (1983)). However, other known bacterial inducible promoters are suitable (Siebenlist *et al.*, *Cell* 20:269 (1980)).

The term "operably linked", as used in reference to a regulatory sequence and a structural nucleotide sequence, means that the regulatory sequence causes regulated expression of the operably linked structural nucleotide sequence. "Expression" refers to the transcription and stable accumulation of sense or antisense RNA derived from the nucleic acid of the present invention. Expression may also refer to translation of mRNA into a polypeptide or protein. "Sense" RNA refers to RNA transcript that includes the mRNA and so can be translated into protein by the cell. "Antisense RNA" refers to a RNA transcript that is complementary to all or part of a target primary transcript or mRNA and that blocks the expression of a target gene (U.S. Pat. No. 5,107,065, incorporated herein by reference). The complementarity of an antisense RNA may be with any part of the specific gene transcript, i.e., at the 5' non-coding sequence, 3' non-translated sequence, introns, or the coding sequence. "RNA transcript" refers to the product resulting from RNA polymerase-catalyzed transcription of a DNA sequence. When the RNA transcript is a perfect complementary copy of the DNA sequence, it is referred to as the primary transcript or it may be a RNA sequence derived from post-transcriptional processing of the primary transcript and is referred to as the mature RNA.

"Regulatory sequences" or "control elements" refer to nucleotide sequences located upstream (5' non-coding sequences), within, or downstream (3' non-translated sequences) of a structural nucleotide sequence, and which influence the transcription, RNA processing or stability, or translation of the associated structural nucleotide sequence. Regulatory sequences may include promoters, translation leader sequences, introns, and polyadenylation recognition sequences.

The promoter sequence may consist of proximal and more distal upstream elements, the latter elements often referred to as enhancers. Accordingly, an "enhancer" is a DNA sequence which can stimulate promoter activity and may be an innate element of the promoter or a heterologous element inserted to enhance the level or tissue-specificity of a promoter. Promoters may be derived in their entirety from a native gene, or be composed of different elements derived from different promoters found in nature, or even comprise synthetic DNA segments. It is understood by those skilled in the art that different promoters may direct the expression of a gene in different tissues or cell types, or at different stages of development, or in response to different environmental conditions.

Promoters for use in bacterial systems also generally contain a Shine-Dalgarno (S.D.) sequence or a consensus sequence thereof operably linked to the DNA encoding the polypeptide of interest. The promoter can be removed from the bacterial source DNA by restriction enzyme digestion and inserted into the vector containing the desired DNA coding sequence, or vice versa.

Alternatively, the expression constructs can be integrated into the bacterial genome with an integrating vector. Integrating vectors typically contain at least one sequence homologous to the bacterial chromosome that allows the vector to integrate. Integrations appear to result from recombinations between homologous DNA in the vector and the bacterial chromosome. For example, integrating vectors constructed with DNA from various *Bacillus* strains integrate into the *Bacillus* chromosome (E.P.O. Pub. No. 127,328). Integrating vectors may also be comprised of bacteriophage or transposon sequences.

Construction of suitable vectors containing one or more of the above-listed components employs standard recombinant DNA techniques. Isolated plasmids or DNA fragments are cleaved, tailored, and re-ligated in the form desired to generate the plasmids required. Examples of available bacterial expression vectors include, but are not limited to, the multifunctional *E. coli* cloning and expression vectors such as Bluescript™ (Stratagene, La Jolla, CA), in which, for example, a *B. thuringiensis* protein or polypeptide of the present invention, may be ligated into the vector in frame with sequences for the amino-terminal Met and the subsequent 7 residues of  $\beta$ -galactosidase so that a hybrid protein is produced; pIN vectors (Van Heeke and Schuster *J. Biol. Chem.* 264:5503-5509 (1989)); and the like. pGEX vectors (Promega, Madison Wis.) may also be used to express foreign polypeptides as fusion proteins with glutathione S-transferase (GST). In general, such fusion proteins are soluble and can easily

be purified from lysed cells by adsorption to glutathione-agarose beads followed by elution in the presence of free glutathione. Proteins made in such systems are designed to include heparin, thrombin or factor XA protease cleavage sites so that the cloned polypeptide of interest can be released from the GST moiety at will.

5 It is, of course, necessary to select the appropriate bacteria taking into consideration replicability of the replicon in the cells of a bacterium. For example, *E. coli*, *Serratia*, or *Salmonella* species can be suitably used as the host when well known plasmids such as pBR322, pBR325, pACYC177, or pKN410 are used to supply the replicon. *E. coli* strain W3110 is a preferred host or parent host because it is a common host strain for recombinant DNA product fermentations. Preferably, the host cell should secrete  
10 minimal amounts of proteolytic enzymes.

Expression vectors for *B. thuringiensis* and related bacilli contain an origin of DNA replication and associated replication factors and a selectable marker gene that will be expressed in *Bacillus* and allow selection of transformed cells. These *Bacillus*-specific plasmid functions may also be present on an *E. coli* or other broad host range cloning vector, forming a plasmid that can replicate and be selected in  
15 either *E. coli* or *B. thuringiensis*, and thus comprising a shuttle vector. Such vectors are well known in the art and usually also contain a *B. thuringiensis*  $\delta^E$  promoter adjacent to one or more convenient cloning sites to allow cloning and expression of exogenous genes during sporulation. Such a promoter is typically found associated with many  $\delta$ -endotoxin genes.

Host cells are transfected and preferably transformed with the above-described vectors and  
20 cultured in conventional nutrient media modified as appropriate for inducing promoters, selecting transformants, or amplifying the genes encoding the desired sequences.

Numerous methods of transfection are known to the ordinarily skilled artisan, for example, calcium phosphate and electroporation. Depending on the host cell used, transformation is done using standard techniques appropriate to such cells. The calcium treatment employing calcium chloride, as  
25 described in section 1.82 of Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, New York: Cold Spring Harbor Laboratory Press, (1989), is generally used for bacterial cells that contain substantial cell-wall barriers. Another method for transformation employs polyethylene glycol/DMSO, as described in Chung and Miller (Chung and Miller, *Nucleic Acids Res.* 16:3580 (1988)). Yet another method is the use of the technique termed electroporation. In addition, bacterial cells can be readily transformed using  
30 various forms of phages (i.e., transducing, temperate, lytic and lysogenic), suicide vectors for inserting DNA directly into the chromosome, and through homologous recombination using either phages, suicide vectors or linear DNA.

Bacterial cells used to produce the polypeptide of interest for purposes of this invention are cultured in suitable media in which the promoters for the nucleic acid encoding the heterologous polypeptide can be artificially induced as described generally, e.g., in Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, New York: Cold Spring Harbor Laboratory Press, (1989). Examples of suitable media are given in U.S. Pat. Nos. 5,304,472 and 5,342,763.

A yeast recombinant construct can typically include one or more of the following: a promoter sequence, fusion partner sequence, leader sequence, transcription termination sequence, a selectable marker. These elements can be combined into an expression cassette, which may be maintained in a replicon, such as an extrachromosomal element (e.g., plasmids) capable of stable maintenance in a host, such as yeast or bacteria. The replicon may have two replication systems, thus allowing it to be maintained, for example, in yeast for expression and in a procaryotic host for cloning and amplification. Examples of such yeast-bacteria shuttle vectors include YEp24 (Botstein *et al.*, *Gene*, 8:17-24 (1979)), pCI/1 (Brake *et al.*, *Proc. Natl. Acad. Sci. USA*, 81:4642-4646 (1984)), and YRp17 (Stinchcomb *et al.*, *J. Mol. Biol.*, 158:157 (1982)). In addition, a replicon may be either a high or low copy number plasmid. A high copy number plasmid will generally have a copy number ranging from about 5 to about 200, and typically about 10 to about 150. A host containing a high copy number plasmid will preferably have at least about 10, and more preferably at least about 20.

Useful yeast promoter sequences can be derived from genes encoding enzymes in the metabolic pathway. Examples of such genes include alcohol dehydrogenase (ADH) (E.P.O. Pub. No. 284044), enolase, glucokinase, glucose-6-phosphate isomerase, glyceraldehyde-3-phosphate-dehydrogenase (GAP or GAPDH), hexokinase, phosphofructokinase, 3-phosphoglycerate mutase, and pyruvate kinase (PyK) (E.P.O. Pub. No. 329203). The yeast PHO5 gene, encoding acid phosphatase, also provides useful promoter sequences (Myanohara *et al.*, *Proc. Natl. Acad. Sci. USA*, 80:1 (1983)). In addition, synthetic promoters which do not occur in nature also function as yeast promoters. Examples of such hybrid promoters include the ADH regulatory sequence linked to the GAP transcription activation region (U.S. Pat. Nos. 4,876,197 and 4,880,734). Other examples of hybrid promoters include promoters which consist of the regulatory sequences of either the ADH2, GAL4, GAL10, or PHO5 genes, combined with the transcriptional activation region of a glycolytic enzyme gene such as GAP or PyK (E.P.O. Pub. No. 164556). Furthermore, a yeast promoter can include naturally occurring promoters of non-yeast origin that have the ability to bind yeast RNA polymerase and initiate transcription. Examples of such promoters include, inter alia, (Cohen *et al.*, *Proc. Natl. Acad. Sci. USA*, 77:1078 (1980); Henikoff *et al.*, *Nature* 283:835 (1981); Hollenberg *et al.*, *Curr. Topics Microbiol. Immunol.*, 96:119 (1981); Mercerau-Puigalon *et al.*, *Gene*, 11:163 (1980); and Panthier *et al.*, *Curr. Genet.*, 2:109 (1980)).

Intracellularly expressed fusion proteins provide an alternative to direct expression of the polypeptides of interest. Typically, a DNA sequence encoding the N-terminal portion of a stable protein, a fusion partner, is fused to the 5' end of heterologous structural nucleotide sequence encoding the desired polypeptide. Upon expression, this construct will provide a fusion of the two amino acid sequences. The DNA sequence at the junction of the two amino acid sequences may or may not encode a cleavable site. See, e.g., E.P.O. Pub. No. 196056. Another example is a ubiquitin fusion protein. Such a ubiquitin fusion protein preferably retains a site for a processing enzyme (e.g. ubiquitin-specific processing protease) to cleave the ubiquitin from the polypeptide of the present invention. Through this method, therefore, a mature polypeptide can be isolated [see, P.C.T. WO 88/024066].

Alternatively, polypeptides or proteins can also be secreted from the cell into the growth media by creating chimeric DNA molecules that encode a fusion protein comprised of a leader sequence fragment that provides for secretion in yeast of the polypeptides. Preferably, there are processing sites encoded between the leader fragment and the polypeptide-encoding sequence fragment that can be cleaved either in vivo or in vitro. The leader sequence fragment typically encodes a signal peptide comprised of hydrophobic amino acids which direct the secretion of the protein from the cell.

DNA encoding suitable signal sequences can be derived from genes for secreted yeast proteins, such as the yeast invertase gene (E.P.O. Pub. No. 12873; J.P.O. Pub. No. 62,096,086) and the A-factor gene (U.S. Pat. No. 4,588,684). Alternatively, leaders of non-yeast origin, such as an interferon leader, exist that also provide for secretion in yeast (E.P.O. Pub. No. 60057).

A preferred class of secretion leaders are those that employ a fragment of the yeast alpha-factor gene, which contains both a "pre" signal sequence, and a "pro" region. The types of alpha-factor fragments that can be employed include the full-length pre-pro alpha factor leader (about 83 amino acid residues) as well as truncated alpha-factor leaders (typically about 25 to about 50 amino acid residues) (U.S. Pat. Nos. 4,546,083 and 4,870,008; and E.P.O. Pub. No. 324274). Additional leaders employing an alpha-factor leader fragment that provides for secretion include hybrid alpha-factor leaders made with a pre-sequence of a first yeast, but a pro-region from a second yeast alpha factor. See, e.g., P.C.T. WO 89/02463.

Examples of transcription terminator sequence and other yeast-recognized termination sequences, such as those coding for glycolytic enzymes, are known to those of skill in the art.

Alternatively, the expression constructs can be integrated into the yeast genome with an integrating vector. Integrating vectors typically contain at least one sequence homologous to a yeast chromosome that allows the vector to integrate, and preferably contain two homologous sequences flanking the expression construct. Integrations appear to result from recombinations between homologous



DNA in the vector and the yeast chromosome (Orr-Weaver et al., Methods in Enzymol., 101:228-245 (1983)). An integrating vector may be directed to a specific locus in yeast by selecting the appropriate homologous sequence for inclusion in the vector. See Orr-Weaver et al., supra. One or more expression constructs may integrate, possibly affecting levels of recombinant protein produced (Rine et al., Proc. Natl. Acad. Sci. USA, 80:6750 (1983)). The chromosomal sequences included in the vector can occur either as a single segment in the vector, which results in the integration of the entire vector, or as two segments homologous to adjacent segments in the chromosome and flanking the expression construct in the vector, which results in the stable integration of only the expression construct.

Expression and transformation vectors, either extrachromosomal replicons or integrating vectors, have been developed for transformation into many yeasts. For example, expression vectors have been developed for, inter alia, the following yeasts: *Candida albicans* (Kurtz, et al., Mol. Cell. Biol., 6:142 (1986)), *Candida maltosa* (Kunze et al., J. Basic Microbiol., 25:141 (1985)); *Hansenula polymorpha* (Gleeson et al., J. Gen. Microbiol. 132:3459 (1986); Roggenkamp et al., Mol. Gen. Genet. 202:302 (1986)); *Kluyveromyces fragilis* (Das et al., J. Bacteriol. 158:1165 (1984)); *Kluyveromyces lactis* (De Louvencourt et al., J. Bacteriol. 154:737 (1983); Van den Berg et al., Bio/Technology 8:135 (1990)); *Pichia guilliermondii* (Kunze et al., J. Basic Microbiol. 25:141 (1985)); *Pichia pastoris* (Cregg et al., Mol. Cell. Biol. 5:3376 (1985); U.S. Pat. Nos. 4,837,148 and 4,929,555); *Saccharomyces cerevisiae* (Hinnen et al., Proc. Natl. Acad. Sci. USA 75:1929 (1978); Ito et al., J. Bacteriol. 153:163 (1983)); *Schizosaccharomyces pombe* (Beach and Nurse, Nature 300:706 (1981)); and *Yarrowia lipolytica* (Davidow, et al., Curr. Genet. 10:380471 (1985); and Gaillardin et al., Curr. Genet. 10:49 (1985)).

Methods of introducing exogenous nucleic acids into yeast hosts are well-known in the art, and typically include either the transformation of spheroplasts or of intact yeast cells treated with alkali cations. Transformation procedures usually vary with the yeast species to be transformed. See e.g., Kurtz et al., Mol. Cell. Biol. 6:142 (1986); Kunze et al., J. Basic Microbiol. 25:141 (1985) for *Candida*. See, e.g., Gleeson et al., J. Gen. Microbiol. 132:3459 (1986); Roggenkamp et al., Mol. Gen. Genet. 202:302 (1986) for *Hansenula*. See, e.g., Das et al., J. Bacteriol. 158:1165 (1984); De Louvencourt et al., J. Bacteriol. 154:1165 (1983); Van den Berg et al., Bio/Technology 8:135 (1990) for *Kluyveromyces*. See, e.g., Cregg et al., Mol. Cell. Biol. 5:3376 (1985); Kunze et al., J. Basic Microbiol. 25:141 (1985); U.S. Pat. Nos. 4,837,148 and 4,929,555 for *Pichia*. See, e.g., Hinnen et al., Proc. Natl. Acad. Sci. USA 75:1929 (1978); Ito et al., J. Bacteriol. 153:163 (1983) for *Saccharomyces*. See, e.g., Beach and Nurse, Nature 300:706 (1981) for *Schizosaccharomyces*. See, e.g., Davidow et al., Curr. Genet. 10:39 (1985); Gaillardin et al., Curr. Genet. 10:49 (1985) for *Yarrowia*.

In order to obtain expression polypeptides or proteins of interest, recombinant microbial host cells derived from the transformants are incubated under conditions which allow expression of the recombinant polypeptide-encoding sequence. These conditions will vary, dependent upon the host cell selected. However, the conditions are readily ascertainable to those of ordinary skill and knowledge in the art.

Detection of polypeptides expressed in the transformed host cell may be performed by several methods. For example, a polypeptide or protein may be detected by its immunological reactivity with antibodies.

Polypeptides or proteins of the present invention may be isolated from the cell by lysis, if formed intracellularly, or isolated from the culture medium, if secreted, by conventional methods.

### Plant Constructs and Plant Transformants

Nucleic acid molecules of the present invention may be used in plant transformation or transfection. Exogenous genetic material may be transferred into a plant cell and the plant cell regenerated into a whole, fertile or sterile plant. Such genetic material may be transferred into either monocotyledons and dicotyledons including but not limited to the plants, alfalfa, *Arabidopsis*, barley, *Brassica*, broccoli, cabbage, citrus, cotton, garlic, oat, oilseed rape, onion, canola, flax, maize, an ornamental plant, pea, peanut, pepper, potato, rice, rye, sorghum, soybean, strawberry, sugarcane, sugarbeet, tomato, wheat, poplar, pine, fir, eucalyptus, apple, lettuce, lentils, grape, banana, tea, turf grasses, sunflower, oil palm, *Phaseolus* etc. Particularly preferred plants would include, *Arabidopsis*, barley, cotton, oat, oilseed rape, rice, maize, soybean, canola, ornamentals, sugarcane, sugarbeet, tomato, potato, wheat and turf grasses (See specifically, Chistou, *Particle Bombardment for Genetic Engineering of Plants*, Biotechnology Intelligence Unit, Academic Press, San Diego, CA (1996) and generally Chistou, *Particle Bombardment for Genetic Engineering of Plants*, Biotechnology Intelligence Unit, Academic Press, San Diego, CA (1996)).

The present invention also relates to a transformed plant cell or plant comprising in its genome an exogenous nucleic acid encoding one or more *B. thuringiensis* proteins or polypeptides of the present invention. The present invention also relates to methods for creating a transgenic plant in which one or more *B. thuringiensis* proteins or polypeptides of the present invention are overexpressed.

As used herein, the term exogenous genetic material means any genetic material, whether naturally occurring or otherwise, from any source that is capable of being inserted into any organism. In a preferred embodiment the exogenous genetic material includes an isolated nucleic acid molecule or fragment thereof of the present invention. In another preferred embodiment of the present invention

exogenous genetic material includes a regulatory element, promoter or partial promoter of the present invention. In a more preferred embodiment of the present invention exogenous genetic material includes a regulatory element, promoter or partial promoter of the present invention and a nucleic acid molecule encoding a *B. thuringiensis* protein or fragments thereof. In a further more preferred embodiment of the present invention exogenous genetic material includes a regulatory element, promoter or partial promoter of the present invention and a nucleic acid molecule of the present invention having a sequence encoding a polypeptide or protein or fragment thereof set forth in Table 1.

One skilled in the art understands that an exogenous nucleic acid can be a heterologous nucleic acid derived from a different species than the species into which the nucleic acid is introduced or can be a nucleic acid derived from the same species as the species into which it is introduced.

The term "overexpression" refers to the expression of a polypeptide or protein encoded by an exogenous nucleic acid introduced into a host cell, wherein said polypeptide or protein is either not normally present in the host cell, or wherein said polypeptide or protein is present in said host cell at a higher level than that normally expressed from the endogenous gene encoding said polypeptide or protein. By "endogenous gene" refers to a native gene in its natural location in the genome of an organism.

The term "genome" as it applies to plant cells encompasses not only chromosomal DNA found within the nucleus, but organelle DNA found within subcellular components of the cell. DNAs of the present invention introduced into plant cells can therefore be either chromosomally integrated or organelle-localized. The term "genome" as it applies to bacteria encompasses both the chromosome and plasmids within a bacterial host cell. Encoding DNAs of the present invention introduced into bacterial host cells can therefore be either chromosomally integrated or plasmid-localized.

Transfer of a nucleic acid molecule that encodes for a protein can result in overexpression of that protein in a transformed cell or transgenic plant. One or more of the proteins or fragments thereof encoded by nucleic acid molecules of the present invention may be overexpressed in a transformed cell or transformed plant. Such overexpression may be the result of transient or stable transfer of the exogenous material. In a preferred embodiment of the present invention, one or more of the *B. thuringiensis* proteins or fragments thereof are overexpressed in a transformed cell or transgenic plant.

Exogenous genetic material may be transferred into a plant cell by the use of a DNA vector or construct designed for such a purpose. Vectors have been engineered for transformation of large DNA inserts into plant genomes. Binary bacterial artificial chromosomes have been designed to replicate in both *E. coli* and *A. tumefaciens* and have all of the features required for transferring large inserts of DNA into plant chromosomes Choi and Wing, <http://genome.clemson.edu/protocols2-nj.html>, July, 1998.

ApBACwch system has been developed to achieve site-directed integration of DNA into the genome. A 150 kb cotton BAC DNA is reported to have been transferred into a specific *lox* site in tobacco by biolistic bombardment and *Cre-lox* site specific recombination.

A construct or vector may also include a plant promoter to express the protein or protein fragment of choice. A number of promoters which are active in plant cells have been described in the literature. These include the nopaline synthase (NOS) promoter (Ebert *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 84:5745-5749 (1987)), the octopine synthase (OCS) promoter (which are carried on tumor-inducing plasmids of *Agrobacterium tumefaciens*), the caulimovirus promoters such as the cauliflower mosaic virus (CaMV) 19S promoter (Lawton *et al.*, *Plant Mol. Biol.* 9:315-324 (1987)) and the CaMV 35S promoter (Odell *et al.*, *Nature* 313:810-812 (1985)), the figwort mosaic virus 35S-promoter, the light-inducible promoter from the small subunit of ribulose-1,5-bis-phosphate carboxylase (ssRUBISCO), the Adh promoter (Walker *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 84:6624-6628 (1987)), the sucrose synthase promoter (Yang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 87:4144-4148 (1990)), the R gene complex promoter (Chandler *et al.*, *The Plant Cell* 1:1175-1183 (1989)), and the chlorophyll a/b binding protein gene promoter, etc. These promoters have been used to create DNA constructs which have been expressed in plants; *see, e.g.*, PCT publication WO 84/02913.

Promoters which are known or are found to cause transcription of DNA in plant cells can be used in the present invention. Such promoters may be obtained from a variety of sources such as plants and plant viruses. It is preferred that the particular promoter selected should be capable of causing sufficient expression to result in the production of an effective amount of protein to cause the desired phenotype. In addition to promoters which are known to cause transcription of DNA in plant cells, other promoters may be identified for use in the current invention by screening a plant cDNA library for genes which are selectively or preferably expressed in the target tissues or cells.

For the purpose of expression in source tissues of the plant, such as the leaf, seed, root or stem, it is preferred that the promoters utilized in the present invention have relatively high expression in these specific tissues. For this purpose, one may choose from a number of promoters for genes with tissue- or cell-specific or -enhanced expression. Examples of such promoters reported in the literature include the chloroplast glutamine synthetase GS2 promoter from pea (Edwards *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 87:3459-3463 (1990)), the chloroplast fructose-1,6-biphosphatase (FBPase) promoter from wheat (Lloyd *et al.*, *Mol. Gen. Genet.* 225:209-216 (1991)), the nuclear photosynthetic ST-LS1 promoter from potato (Stockhaus *et al.*, *EMBO J.* 8:2445-2451 (1989)), the phenylalanine ammonia-lyase (PAL) promoter and the chalcone synthase (CHS) promoter from *B. thuringiensis thaliana*. Also reported to be active in photosynthetically active tissues are the ribulose-1,5-bisphosphate carboxylase (RbcS) promoter from

eastern larch (*Larix laricina*), the promoter for the *cab* gene, *cab6*, from pine (Yamamoto *et al.*, *Plant Cell Physiol.* 35:773-778 (1994)), the promoter for the *Cab-1* gene from wheat (Fejes *et al.*, *Plant Mol. Biol.* 15:921-932 (1990)), the promoter for the *CAB-1* gene from spinach (Lubberstedt *et al.*, *Plant Physiol.* 104:97-1006 (1994)), the promoter for the *cab1R* gene from rice (Luan *et al.*, *Plant Cell.* 4:971-981 (1992)), the pyruvate, orthophosphate dikinase (PPDK) promoter from maize (Matsuoka *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 90:9586-9590 (1993)), the promoter for the tobacco *Lhcb1\*2* gene (Cerdan *et al.*, *Plant Mol. Biol.* 33:245-255. (1997)), the *A. nidulans thaliana* SUC2 sucrose-H<sup>+</sup> symporter promoter (Truernit *et al.*, *Planta.* 196:564-570 (1995)), and the *th spergillus thaliana* SUC2 sucrose-H<sup>+</sup> symporter promoter (Truernit *et al.*, *Planta.* 196:564-570 (1995)), and the promoter for the thylakoid membrane proteins from spinach (*psaD*, *psaF*, *psaE*, *PC*, *FNR*, *atpC*, *atpD*, *cab*, *rbcS*). Other promoters for the chlorophyll a/b-binding proteins may also be utilized in the present invention, such as the promoters for *Lhcb* gene and *PsbP* gene from white mustard (*Sinapis alba*; Kretsch *et al.*, *Plant Mol. Biol.* 28:219-229 (1995)).

For the purpose of expression in sink tissues of the plant, such as the tuber of the potato plant, the fruit of tomato, or the seed of maize, wheat, rice, and barley, it is preferred that the promoters utilized in the present invention have relatively high expression in these specific tissues. A number of promoters for genes with tuber-specific or -enhanced expression are known, including the class I patatin promoter (Bevan *et al.*, *EMBO J.* 8:1899-1906 (1986); Jefferson *et al.*, *Plant Mol. Biol.* 14:995-1006 (1990)), the promoter for the potato tuber ADPGPP genes, both the large and small subunits, the sucrose synthase promoter (Salanoubat and Belliard, *Gene.* 60:47-56 (1987), Salanoubat and Belliard, *Gene.* 84:181-185 (1989)), the promoter for the major tuber proteins including the 22 kd protein complexes and proteinase inhibitors (Hannapel, *Plant Physiol.* 101:703-704 (1993)), the promoter for the granule bound starch synthase gene (GBSS) (Visser *et al.*, *Plant Mol. Biol.* 17:691-699 (1991)), and other class I and II patatins promoters (Koster-Topfer *et al.*, *Mol. Gen. Genet.* 219:390-396 (1989); Mignery *et al.*, *Gene.* 62:27-44 (1988)).

Other promoters can also be used to express a fructose 1,6 bisphosphate aldolase gene in specific tissues, such as seeds or fruits. The promoter for  $\beta$ -conglycinin (Chen *et al.*, *Dev. Genet.* 10:112-122 (1989)) or other seed-specific promoters such as the napin and phaseolin promoters, can be used. The zeins are a group of storage proteins found in maize endosperm. Genomic clones for zein genes have been isolated (Pedersen *et al.*, *Cell* 29:1015-1026 (1982)), and the promoters from these clones, including the 15 kD, 16 kD, 19 kD, 22 kD, 27 kD, and gamma genes, could also be used. Other promoters known to function, for example, in maize, include the promoters for the following genes: *waxy*, *Brittle*, *Shrunken 2*, Branching enzymes I and II, starch synthases, debranching enzymes, oleosins, glutelins, and sucrose

synthases. A particularly preferred promoter for maize endosperm expression is the promoter for the glutelin gene from rice, more particularly the Osgt-1 promoter (Zheng *et al.*, *Mol. Cell Biol.* 13:5829-5842 (1993)). Examples of promoters suitable for expression in wheat include those promoters for the ADPglucose pyrophosphorylase (ADPGPP) subunits, the granule bound and other starch synthases, the branching and debranching enzymes, the embryogenesis-abundant proteins, the gliadins, and the glutenins. Examples of such promoters in rice include those promoters for the ADPGPP subunits, the granule bound and other starch synthases, the branching enzymes, the debranching enzymes, sucrose synthases, and the glutelins. A particularly preferred promoter is the promoter for rice glutelin, Osgt-1. Examples of such promoters for barley include those for the ADPGPP subunits, the granule bound and other starch synthases, the branching enzymes, the debranching enzymes, sucrose synthases, the hordeins, the embryo globulins, and the aleurone specific proteins.

Root specific promoters may also be used. An example of such a promoter is the promoter for the acid chitinase gene (Samac *et al.*, *Plant Mol. Biol.* 25:587-596 (1994)). Expression in root tissue could also be accomplished by utilizing the root specific subdomains of the CaMV35S promoter that have been identified (Lam *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 86:7890-7894 (1989)). Other root cell specific promoters include those reported by Conkling *et al.* (Conkling *et al.*, *Plant Physiol.* 93:1203-1211 (1990)).

Additional promoters that may be utilized are described, for example, in U.S. Patent Nos. 5,378,619; 5,391,725; 5,428,147; 5,447,858; 5,608,144; 5,608,144; 5,614,399; 5,633,441; 5,633,435 and 4,633,436, all of which are herein incorporated in their entirety. In addition, a tissue specific enhancer may be used (Fromm *et al.*, *The Plant Cell* 1:977-984 (1989)). It is further understood that one or more of the promoters of the present invention may be used.

Constructs or vectors may also include, with the coding region of interest, a nucleic acid sequence that acts, in whole or in part, to terminate transcription of that region. For example, such sequences have been isolated including the Tr7 3' sequence and the nos 3' sequence (Ingelbrecht *et al.*, *The Plant Cell* 1:671-680 (1989); Bevan *et al.*, *Nucleic Acids Res.* 11:369-385 (1983)), or the like. It is understood that one or more sequences of the present invention that act, to terminate transcription may be used.

A vector or construct may also include other regulatory elements. Examples of such include the Adh intron 1 (Callis *et al.*, *Genes and Develop.* 1:1183-1200 (1987)), the sucrose synthase intron (Vasil *et al.*, *Plant Physiol.* 91:1575-1579 (1989)) and the TMV omega element (Gallie *et al.*, *The Plant Cell* 1:301-311 (1989)). These and other regulatory elements may be included when appropriate. It is also understood that one or more of the regulatory regions of the present invention may be used.

A vector or construct may also include a selectable marker. Selectable markers may also be used to select for plants or plant cells that contain the exogenous genetic material. Examples of such include, but are not limited to, a neo gene (Potrykus *et al.*, *Mol. Gen. Genet.* 199:183-188 (1985)) which codes for kanamycin resistance and can be selected for using kanamycin, G418, etc.; a bar gene which codes for bialaphos resistance; a mutant EPSP synthase gene (Hinchee *et al.*, *Bio/Technology* 6:915-922 (1988)) which encodes glyphosate resistance; a nitrilase gene which confers resistance to bromoxynil (Stalker *et al.*, *J. Biol. Chem.* 263:6310-6314 (1988)); a mutant acetolactate synthase gene (ALS) which confers imidazolinone or sulphonylurea resistance (European Patent Application 154,204 (Sept. 11, 1985)); and a methotrexate resistant DHFR gene (Thillet *et al.*, *J. Biol. Chem.* 263:12500-12508 (1988)).

A vector or construct may also include a transit peptide. Incorporation of a suitable chloroplast transit peptide may also be employed (European Patent Application Publication Number 0218571). Translational enhancers may also be incorporated as part of the vector DNA. DNA constructs could contain one or more 5' non-translated leader sequences which may serve to enhance expression of the gene products from the resulting mRNA transcripts. Such sequences may be derived from the promoter selected to express the gene or can be specifically modified to increase translation of the mRNA. Such regions may also be obtained from viral RNAs, from suitable eukaryotic genes, or from a synthetic gene sequence. For a review of optimizing expression of transgenes, see Koziel *et al.*, *Plant Mol. Biol.* 32:393-405 (1996).

A vector or construct may also include a screenable marker. Screenable markers may be used to monitor expression. Exemplary screenable markers include a  $\beta$ -glucuronidase or uidA gene (GUS) which encodes an enzyme for which various chromogenic substrates are known (Jefferson, *Plant Mol. Biol. Rep.* 5:387-405 (1987); Jefferson *et al.*, *EMBO J.* 6:3901-3907 (1987)); an R-locus gene, which encodes a product that regulates the production of anthocyanin pigments (red color) in plant tissues (Dellaporta *et al.*, *Stadler Symposium* 11:263-282 (1988)); a  $\beta$ -lactamase gene (Sutcliffe *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 75:3737-3741 (1978)), a gene which encodes an enzyme for which various chromogenic substrates are known (e.g., PADAC, a chromogenic cephalosporin); a luciferase gene (Ow *et al.*, *Science* 234:856-859 (1986)) a xyle gene (Zukowsky *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 80:1101-1105 (1983)) which encodes a catechol dioxygenase that can convert chromogenic catechols; an  $\alpha$ -amylase gene (Ikata *et al.*, *Bio/Technol.* 8:241-242 (1990)); a tyrosinase gene (Katz *et al.*, *J. Gen. Microbiol.* 129:2703-2714 (1983)) which encodes an enzyme capable of oxidizing tyrosine to DOPA and dopaquinone which in turn condenses to melanin; an  $\alpha$ -galactosidase, which will turn a chromogenic  $\alpha$ -galactose substrate.

Included within the terms "selectable or screenable marker genes" are also genes which encode a secretable marker whose secretion can be detected as a means of identifying or selecting for transformed cells. Examples include markers which encode a secretable antigen that can be identified by antibody interaction, or even secretable enzymes which can be detected catalytically. Secretable proteins fall into a number of classes, including small, diffusible proteins detectable, *e.g.*, by ELISA, small active enzymes detectable in extracellular solution (*e.g.*,  $\alpha$ -amylase,  $\beta$ -lactamase, phosphinothricin transferase), or proteins which are inserted or trapped in the cell wall (such as proteins which include a leader sequence such as that found in the expression unit of extension or tobacco PR-S). Other possible selectable and/or screenable marker genes will be apparent to those of skill in the art.

There are many methods for introducing nucleic acid molecules into plant cells. Suitable methods are believed to include virtually any method by which nucleic acid molecules may be introduced into a cell, such as by *Agrobacterium* infection or direct delivery of nucleic acid molecules such as, for example, by PEG-mediated transformation, by electroporation or by acceleration of DNA coated particles, etc. (Potrykus, *Ann. Rev. Plant Physiol. Plant Mol. Biol.* 42:205-225 (1991); Vasil, *Plant Mol. Biol.* 25:925-937 (1994). For example, electroporation has been used to transform maize protoplasts (Fromm *et al.*, *Nature* 312:791-793 (1986)).

Technology for introduction of DNA into cells is well known to those of skill in the art. Four general methods for delivering a gene into cells have been described: (1) chemical methods (Graham and van der Eb, *Virology* 54:536-539 (1973)); (2) physical methods such as microinjection (Capecchi, *Cell* 22:479-488 (1980), electroporation (Wong and Neumann, *Biochem. Biophys. Res. Commun.* 107:584-587 (1982); Fromm *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 82:5824-5828 (1985); U. S. Patent No. 5,384,253; and the gene gun (Johnston and Tang, *Methods Cell Biol.* 43:353-365 (1994); (3) viral vectors (Clapp, *Clin. Perinatol.* 20:155-168 (1993); Lu *et al.*, *J. Exp. Med.* 178:2089-2096 (1993); Eglitis and Anderson, *Biotechnology* 6:608-614 (1988)); and (4) receptor-mediated mechanisms (Curiel *et al.*, *Hum. Gen. Ther.* 3:147-154 (1992); Wagner *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 89:6099-6103 (1992)).

Acceleration methods that may be used include, for example, microprojectile bombardment and the like. One example of a method for delivering transforming nucleic acid molecules to plant cells is microprojectile bombardment. This method has been reviewed by Yang and Christou, eds., *Particle Bombardment Technology for Gene Transfer*, Oxford Press, Oxford, England (1994)). Non-biological particles (microprojectiles) that may be coated with nucleic acids and delivered into cells by a propelling force. Exemplary particles include those comprised of tungsten, gold, platinum, and the like.

A particular advantage of microprojectile bombardment, in addition to it being an effective means of reproducibly, and stably transforming monocotyledons, is that neither the isolation of protoplasts



(Cristou *et al.*, *Plant Physiol.* 87:671-674 (1988)) nor the susceptibility of *Agrobacterium* infection is required. An illustrative embodiment of a method for delivering DNA into maize cells by acceleration is a biolistics alpha-particle delivery system, which can be used to propel particles coated with DNA through a screen, such as a stainless steel or Nytex screen, onto a filter surface covered with corn cells cultured in suspension. Gordon-Kamm *et al.*, describes the basic procedure for coating tungsten particles with DNA (*Plant Cell* 2:603-618 (1990)). The screen disperses the tungsten nucleic acid particles so that they are not delivered to the recipient cells in large aggregates. A particle delivery system suitable for use with the present invention is the helium acceleration PDS-1000/He gun which is available from Bio-Rad Laboratories (Bio-Rad, Hercules, CA) Sanford *et al.*, *Technique* 3:3-16 (1991)).

For the bombardment, cells in suspension may be concentrated on filters. Filters containing the cells to be bombarded are positioned at an appropriate distance below the microprojectile stopping plate. If desired, one or more screens are also positioned between the gun and the cells to be bombarded.

Alternatively, immature embryos or other target cells may be arranged on solid culture medium. The cells to be bombarded are positioned at an appropriate distance below the macroprojectile stopping plate. If desired, one or more screens are also positioned between the acceleration device and the cells to be bombarded. Through the use of techniques set forth herein one may obtain up to 1000 or more foci of cells transiently expressing a marker gene. The number of cells in a focus which express the exogenous gene product 48 hours post-bombardment often range from one to ten and average one to three.

In another alternative embodiment, plastids can be stably transformed. Methods suitable for plastid transformation in higher plants include particle gun delivery of DNA containing a selectable marker and targeting of the DNA to the plastid genome through homologous recombination (Svab *et al.* *Proc. Natl. Acad. Sci. (U.S.A.)* 87:8526-8530 (1990); Svab and Maliga *Proc. Natl. Acad. Sci. (U.S.A.)* 90:913-917 (1993); Staub and Maliga, *EMBO J.* 12:601-606 (1993), U.S. Patents 5,451,513 and 5,545,818).

In bombardment transformation, one may optimize the prebombardment culturing conditions and the bombardment parameters to yield the maximum numbers of stable transformants. Both the physical and biological parameters for bombardment are important in this technology. Physical factors are those that involve manipulating the DNA/microprojectile precipitate or those that affect the flight and velocity of either the macro- or microprojectiles. Biological factors include all steps involved in manipulation of cells before and immediately after bombardment, the osmotic adjustment of target cells to help alleviate the trauma associated with bombardment, and also the nature of the transforming DNA, such as linearized DNA or intact supercoiled plasmids. It is believed that pre-bombardment manipulations are especially important for successful transformation of immature embryos.

Accordingly, it is contemplated that one may wish to adjust various aspects of the bombardment parameters in small scale studies to fully optimize the conditions. One may particularly wish to adjust physical parameters such as gap distance, flight distance, tissue distance, and helium pressure. One may also minimize the trauma reduction factors by modifying conditions which influence the physiological state of the recipient cells and which may therefore influence transformation and integration efficiencies. For example, the osmotic state, tissue hydration and the subculture stage or cell cycle of the recipient cells may be adjusted for optimum transformation. The execution of other routine adjustments will be known to those of skill in the art in light of the present disclosure.

*Agrobacterium*-mediated transfer is a widely applicable system for introducing genes into plant cells because the DNA can be introduced into whole plant tissues, thereby bypassing the need for regeneration of an intact plant from a protoplast. The use of *Agrobacterium*-mediated plant integrating vectors to introduce DNA into plant cells is well known in the art. See, for example, the methods described (Fraley *et al.*, *Biotechnology* 3:629-635 (1985); Rogers *et al.*, *Meth. Enzymol.* 153:253-277 (1987). Further, the integration of the Ti-DNA is a relatively precise process resulting in few rearrangements. The region of DNA to be transferred is defined by the border sequences, and intervening DNA is usually inserted into the plant genome as described (Spielmann *et al.*, *Mol. Gen. Genet.* 205:34 (1986)).

Modern *Agrobacterium* transformation vectors are capable of replication in *E. coli* as well as *Agrobacterium*, allowing for convenient manipulations as described (Klee *et al.*, *In: Plant DNA Infectious Agents*, Hohn and Schell, eds., Springer-Verlag, New York, pp. 179-203 (1985)). Moreover, recent technological advances in vectors for *Agrobacterium*-mediated gene transfer have improved the arrangement of genes and restriction sites in the vectors to facilitate construction of vectors capable of expressing various polypeptide coding genes. The vectors described have convenient multi-linker regions flanked by a promoter and a polyadenylation site for direct expression of inserted polypeptide coding genes and are suitable for present purposes (Rogers *et al.*, *Methods Enzymol.* 153:253-277 (1987)). In addition, *Agrobacterium* containing both armed and disarmed Ti genes can be used for the transformations. In those plant strains where *Agrobacterium*-mediated transformation is efficient, it is the method of choice because of the facile and defined nature of the gene transfer.

A transgenic plant formed using *Agrobacterium* transformation methods typically contains a single insert on one chromosome. Such transgenic plants can be referred to as being heterozygous for the added gene. More preferred is a transgenic plant that is homozygous for the added structural gene; *i.e.*, a transgenic plant that contains two added genes, one gene at the same locus on each chromosome of a chromosome pair. A homozygous transgenic plant can be obtained by sexually mating (selfing) an

independent segregant transgenic plant that contains a single added gene, germinating some of the seed produced and analyzing the resulting plants produced for the gene of interest.

It is also to be understood that two different transgenic plants can also be mated to produce offspring that contain two independently segregating added, exogenous genes.

5       The present invention also provides for parts of the plants of the present invention. Plant parts, without limitation, include seed, endosperm, ovule and pollen. In a particularly preferred embodiment of the present invention, the plant part is a seed.

Transformation of plant protoplasts can be achieved using methods based on calcium phosphate precipitation, polyethylene glycol treatment, electroporation, and combinations of these treatments. See  
10       for example (Potrykus *et al.*, *Mol. Gen. Genet* 205:193-200 (1986); Lorz *et al.*, *Mol. Gen. Genet.* 199:178, (1985); Fromm *et al.*, *Nature* 319:791,(1986); Uchimiya *et al.*, *Mol. Gen. Genet.* 204:204 (1986); Callis *et al.*, *Genes and Development* 1183 (1987); Marcotte *et al.*, *Nature* 335:454 (1988)).

Application of these systems to different plant strains depends upon the ability to regenerate that particular plant strain from protoplasts. Illustrative methods for the regeneration of cereals from  
15       protoplasts are described (Fujimura *et al.*, *Plant Tissue Culture Letters* 2:74 (1985); Toriyama *et al.*, *Theor Appl. Genet.* 205:34 (1986); Yamada *et al.*, *Plant Cell Rep.* 4:85 (1986); Abdullah *et al.*, *Biotechnology* 4:1087 (1986)).

To transform plant strains that cannot be successfully regenerated from protoplasts, other ways to introduce DNA into intact cells or tissues can be utilized. For example, regeneration of cereals from  
20       immature embryos or explants can be effected as described (Vasil, *Biotechnology* 6: 397 (1988)). In addition, "particle gun" or high-velocity microprojectile technology can be utilized (Vasil *et al.*, *Bio/Technology* 10:667, (1992)).

Using the latter technology, DNA is carried through the cell wall and into the cytoplasm on the surface of small metal particles as described (Klein *et al.*, *Nature* 328:70 (1987); Klein *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 85:8502-8505 (1988); McCabe *et al.*, *Biotechnology* 6 :923 (1988)). The metal  
25       particles penetrate through several layers of cells and thus allow the transformation of cells within tissue explants.

Other methods of cell transformation can also be used and include but are not limited to introduction of DNA into plants by direct DNA transfer into pollen (Hess *et al.*, *Intern Rev. Cytol.*  
30       107:367 (1987); Luo *et al.*, *Plant Mol. Biol.. Reporter* 6:165 (1988)), by direct injection of DNA into reproductive organs of a plant (Pena *et al.*, *Nature* 325:274 (1987)), or by direct injection of DNA into the cells of immature embryos followed by the rehydration of desiccated embryos (Neuhaus *et al.*, *Theor. Appl. Genet.* 75:30 (1987)).

The regeneration, development, and cultivation of plants from single plant protoplast transformants or from various transformed explants is well known in the art (Weissbach and Weissbach, In: *Methods for Plant Molecular Biology*, (Eds.), Academic Press, Inc. San Diego, CA, (1988)). This regeneration and growth process typically includes the steps of selection of transformed cells, culturing those individualized cells through the usual stages of embryonic development through the rooted plantlet stage. Transgenic embryos and seeds are similarly regenerated. The resulting transgenic rooted shoots are thereafter planted in an appropriate plant growth medium such as soil.

The development or regeneration of plants containing the foreign, exogenous gene that encodes a protein of interest is well known in the art. Preferably, the regenerated plants are self-pollinated to provide homozygous transgenic plants, as discussed before. Otherwise, pollen obtained from the regenerated plants is crossed to seed-grown plants of agronomically important lines. Conversely, pollen from plants of these important lines is used to pollinate regenerated plants. A transgenic plant of the present invention containing a desired polypeptide is cultivated using methods well known to one skilled in the art.

There are a variety of methods for the regeneration of plants from plant tissue. The particular method of regeneration will depend on the starting plant tissue and the particular plant species to be regenerated.

Methods for transforming dicots, primarily by use of *Agrobacterium tumefaciens*, and obtaining transgenic plants have been published for cotton (U. S. Patent No. 5,004,863; U.S. Patent No. 5,159,135; U.S. Patent No. 5,518,908); soybean (U. S. Patent No. 5,569,834; U. S. Patent No. 5,416,011; McCabe *et al.*, *Biotechnology* 6:923 (1988), Christou *et al.*, *Plant Physiol.* 87:671-674 (1988)); *Brassica* ( U. S. Patent No. 5,463,174); peanut (Cheng *et al.*, *Plant Cell Rep.* 15:653-657 (1996), McKently *et al.*, *Plant Cell Rep.* 14:699-703 (1995)); papaya (Yang *et al.*, (1996)); pea (Grant *et al.*, *Plant Cell Rep.* 15:254-258, (1995)).

Transformation of monocotyledons using electroporation, particle bombardment, and *Agrobacterium* have also been reported. Transformation and plant regeneration have been achieved in asparagus (Bytebier *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 84:5345 (1987)); barley (Wan and Lemaux, *Plant Physiol* 104:37 (1994)); maize (Rhodes *et al.*, *Science* 240:204 (1988), Gordon-Kamm *et al.*, *Plant Cell* 2:603 (1990), Fromm *et al.*, *Bio/Technology* 8:833 (1990), Koziel *et al.*, *Bio/Technology* 11:194 (1993), Armstrong *et al.*, *Crop Science* 35:550-557 (1995)); oat (Somers *et al.*, *Bio/Technology* 10:1589 (1992)); orchardgrass (Horn *et al.*, *Plant Cell Rep.* 7:469 (1988)); rice (Toriyama *et al.*, *Theor Appl. Genet.* 205:34 (1986); Park *et al.*, *Plant Mol. Biol.* 32:1135-1148 (1996); Abedinia *et al.*, *Aust. J. Plant Physiol.* 24:133-141 (1997); Zhang and Wu, *Theor. Appl. Genet.* 76:835 (1988); Zhang *et al.* *Plant Cell*

Rep. 7:379 (1988); Battraw and Hall, *Plant Sci.* 86:191-202 (1992); Christou *et al.*, *Bio/Technology* 9:957 (1991)); sugarcane (Bower and Birch, *Plant J.* 2:409 (1992)); tall fescue (Wang *et al.*, *Bio/Technology* 10:691 (1992)), and wheat (Vasil *et al.*, *Bio/Technology* 10:667 (1992); U. S. Patent No. 5,631,152.

Assays for gene expression based on the transient expression of cloned nucleic acid constructs have been developed by introducing the nucleic acid molecules into plant cells by polyethylene glycol treatment, electroporation, or particle bombardment (Marcotte *et al.*, *Nature* 335:454-457 (1988); Marcotte *et al.*, *Plant Cell* 1:523-532 (1989); McCarty *et al.*, *Cell* 66:895-905 (1991); Hattori *et al.*, *Genes Dev.* 6:609-618 (1992); Goff *et al.*, *EMBO J.* 9:2517-2522 (1990)). Transient expression systems may be used to functionally dissect gene constructs (See generally, Mailga *et al.*, *Methods in Plant Molecular Biology*, Cold Spring Harbor Press (1995)).

Any of the nucleic acid molecules of the present invention may be introduced into a plant cell in a permanent or transient manner in combination with other genetic elements such as vectors, promoters, enhancers etc. Further any of the nucleic acid molecules encoding a *B. thuringiensis* protein or fragment thereof of the present invention may be introduced into a plant cell in a manner that allows for over expression of the protein or fragment thereof encoded by the nucleic acid molecule.

Antibodies have been expressed in plants (Hiatt *et al.*, *Nature* 342:76-78 (1989); Conrad and Fielder, *Plant Mol. Biol.* 26:1023-1030 (1994)). Cytoplasmic expression of a scFv (single-chain Fv antibodies) has been reported to delay infection by artichoke mottled crinkle virus. Transgenic plants that express antibodies directed against endogenous proteins may exhibit a physiological effect (Philips *et al.*, *EMBO J.* 16:4489-4496 (1997); Marion-Poll, *Trends in Plant Science* 2:447-448 (1997)). For example, expressed anti-abscisic antibodies reportedly result in a general perturbation of seed development (Philips *et al.*, *EMBO J.* 16:4489-4496 (1997)).

Antibodies that are catalytic may also be expressed in plants (abzymes). The principle behind abzymes is that since antibodies may be raised against many molecules, this recognition ability can be directed toward generating antibodies that bind transition states to force a chemical reaction forward (Persidas, *Nature Biotechnology* 15:1313-1315 (1997); Baca *et al.*, *Ann. Rev. Biophys. Biomol. Struct.* 26:461-493 (1997)). The catalytic abilities of abzymes may be enhanced by site directed mutagenesis. Examples of abzymes are, for example, set forth in U.S. Patent No: 5,658,753; U.S. Patent No. 5,632,990; U.S. Patent No. 5,631,137; U.S. Patent 5,602,015; U.S. Patent No. 5,559,538; U.S. Patent No. 5,576,174; U.S. Patent No. 5,500,358; U.S. Patent 5,318,897; U.S. Patent No. 5,298,409; U.S. Patent No. 5,258,289 and U.S. Patent No. 5,194,585, all of which are herein incorporated in their entirety.

It is understood that any of the antibodies of the present invention may be expressed in plants and that such expression can result in a physiological effect. It is also understood that any of the expressed antibodies may be catalytic.

## 5 Exemplary Uses of the Agents of the Present Invention

Nucleic acid molecules and fragments thereof of the present invention may be employed to obtain other nucleic acid molecules from the same species. Such nucleic acid molecules include the nucleic acid molecules that encode the complete coding sequence of a protein and promoters and flanking sequences of such molecules. In addition, such nucleic acid molecules include nucleic acid molecules  
10 that encode for other isozymes or gene family members. Such molecules can be readily obtained by using the above-described nucleic acid molecules or fragments thereof to screen genomic libraries obtained from *B. thuringiensis*. Methods for forming such libraries are well known in the art.

Nucleic acid molecules and fragments thereof of the present invention may also be employed to obtain other nucleic acid molecules such as nucleic acid homologues. Such homologues include the  
15 nucleic acid nucleic acid homologues of non-*B. thuringiensis* species including the nucleic acid molecules that encode, in whole or in part, protein homologues of other species or other organisms, sequences of genetic elements such as promoters and transcriptional regulatory elements. Such molecules can be readily obtained by using the above-described nucleic acid molecules or fragments thereof to screen cDNA or genomic libraries. Methods for forming such libraries are well known in the  
20 art. Such homologue molecules may differ in their nucleotide sequences from those found in one or more of SEQ ID NO: 1 through SEQ ID NO: 8283 or complements thereof because complete complementarity is not needed for stable hybridization. The nucleic acid molecules of the present invention therefore also include molecules that, although capable of specifically hybridizing with the nucleic acid molecules may lack "complete complementarity." In a particular embodiment, methods or 3' or 5' RACE may be used  
25 (Frohman, M.A. *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 85:8998-9002 (1988); Ohara, O. *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 86:5673-5677 (1989)) to obtain such sequences.

Any of a variety of methods may be used to obtain one or more of the above-described nucleic acid molecules (Zamechik *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 83:4143-4146 (1986); Goodchild *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 85: 5507-5511 (1988); Wickstrom *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)*  
30 85:1028-1032 (1988); Holt *et al.*, *Molec. Cell. Biol.* 8:963-973 (1988); Gerwartz *et al.*, *Science* 242: 1303-1306 (1988); Anfossi *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 86:3379-3383 (1989); Becker *et al.*, *EMBO J.* 8:3685-3691 (1989)). Automated nucleic acid synthesizers may be employed for this purpose. In lieu of such synthesis, the disclosed nucleic acid molecules may be used to define a pair of primers that can be

used with the polymerase chain reaction (Mullis *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* 51:263-273 (1986); Erlich *et al.*, European Patent 50,424; European Patent 84,796, European Patent 258,017; European Patent 237,362; Mullis, European Patent 201,184; Mullis *et al.*, U.S. Patent 4,683,202; Erlich, U.S. Patent 4,582,788; and Saiki *et al.*, U.S. Patent 4,683,194) to amplify and obtain any desired nucleic acid molecule or fragment.

The nucleic acid molecules of the present invention may be used for physical mapping. Physical mapping, in conjunction with linkage analysis, can enable the isolation of genes. Physical mapping has been reported to identify the markers closest in terms of genetic recombination to a gene target for cloning. Once a DNA marker is linked to a gene of interest, the chromosome walking technique can be used to find the genes via overlapping clones. For chromosome walking, random molecular markers or established molecular linkage maps are used to conduct a search to localize the gene adjacent to one or more markers. A chromosome walk (Bukanov and Berg, *Mo. Microbiol.* 11:509-523 (1994); Birkenbihl and Vielmetter *Nucleic Acids Res.* 17:5057-5069 (1989); Wenzel and Herrmann, *Nucleic Acids Res.* 16:8323-8336 (1988) is then initiated from the closest linked marker. Starting from the selected clones, labeled probes specific for the ends of the insert DNA are synthesized and used as probes in hybridizations against a representative library. Clones hybridizing with one of the probes are picked and serve as templates for the synthesis of new probes; by subsequent analysis, contigs are produced.

The degree of overlap of the hybridizing clones used to produce a contig can be determined by comparative restriction analysis. Comparative restriction analysis can be carried out in different ways all of which exploit the same principle; two clones of a library are very likely to overlap if they contain a limited number of restriction sites for one or more restriction endonucleases located at the same distance from each other. The most frequently used procedures are, fingerprinting (Coulson *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 83:7821-7821, (1986); Knott *et al.*, *Nucleic Acids Res.* 16:2601-2612 (1988); Eiglmeier *et al.*, *Mol. Microbiol.* 7:197-206 (1993), 1993), restriction fragment mapping (Smith and Birnstiel, *Nucleic Acids Res.* 3:2387-2398 (1976)); or the "landmarking" technique (Charlebois *et al.*, *J. Mol. Biol.* 222:509-524 (1991)).

It is understood that the nucleic acid molecules of the present invention may in one embodiment be used in physical mapping. In a preferred embodiment, nucleic acid molecules of the present invention may in one embodiment be used in the physical mapping of *B. thuringiensis*.

Nucleic acid molecules of the present invention can be used in comparative mapping. Comparative mapping within families provides a method to assess the degree of sequence conservation, gene order, ploidy of species, ancestral relationships and the rates at which individual genomes are evolving. Comparative mapping has been carried out by cross-hybridizing molecular markers across

species within a given family. As in genetic mapping, molecular markers are needed but instead of direct hybridization to mapping filters, the markers are used to select large insert clones from a total genomic DNA library of a related species. The selected clones, each a representative of a single marker, can then be used to physically map the region in the target species. The advantage of this method for comparative mapping is that no mapping population or linkage map of the target species is needed and the clones may also be used in other closely related species. By comparing the results obtained by genetic mapping in model organisms, with those from other species, similarities of genomic structure among species can be established. Cross-hybridization of RFLP markers has been reported and conserved gene order has been established in many studies. Such macroscopic synteny is utilized for the estimation of correspondence of loci among these organisms. It is understood that markers of the present invention may in another embodiment be used in comparative mapping. In a preferred embodiment the markers of present invention may be used in the comparative mapping of spore-forming Gram-positive bacteria.

In an aspect of the present invention, one or more of the agents of the present invention may be used to detecting the presence, absence or level of an organism, preferably a *Bacillus*, and more preferably a *B. thuringiensis* in a sample. In another aspect of the present invention, one or more of the nucleic molecules of the present invention are used to determine the level (i.e., the concentration of mRNA in a sample, etc.) or pattern (i.e., the kinetics of expression, rate of decomposition, stability profile, etc.) of the expression encoded in part or whole by one or more of the nucleic acid molecule of the present invention (collectively, the "Expression Response" of a cell or tissue). As used herein, the Expression Response manifested by a cell or tissue is said to be "altered" if it differs from the Expression Response of cells or tissues of organisms not exhibiting the phenotype. To determine whether a Expression Response is altered, the Expression Response manifested by the cell or tissue of the organism exhibiting the phenotype is compared with that of a similar cell or tissue sample of a organism not exhibiting the phenotype. As will be appreciated, it is not necessary to re-determine the Expression Response of the cell or tissue sample of organisms not exhibiting the phenotype each time such a comparison is made; rather, the Expression Response of a particular organism may be compared with previously obtained values of normal organism. As used herein, the phenotype of the organism is any of one or more characteristics of an organism.

Nucleic acid molecules of the present invention can be used to monitor expression. A microarray-based method for high-throughput monitoring of gene expression may be utilized to measure gene-specific hybridization targets. This 'chip'-based approach involves using microarrays of nucleic acid molecules as gene-specific hybridization targets to quantitatively measure expression of the corresponding genes (Schena *et al.*, *Science* 270:467-470 (1995); Shalon, Ph.D. Thesis, Stanford



University (1996)). Every nucleotide in a large sequence can be queried at the same time. Hybridization can be used to efficiently analyze nucleotide sequences.

Several microarray methods have been described. One method compares the sequences to be analyzed by hybridization to a set of oligonucleotides or cDNA molecules representing all possible subsequences (Bains and Smith, *J. Theor. Biol.* 135:303 (1989)). A second method hybridizes the sample to an array of oligonucleotide or cDNA probes. An array consisting of oligonucleotides or cDNA molecules complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect differences between the target and a reference sequence. Nucleic acid molecules microarrays may also be screened with protein molecules or fragments thereof to determine nucleic acid molecules that specifically bind protein molecules or fragments thereof.

The microarray approach may also be used with polypeptide targets (U.S. Patent No. 5,445,934; U.S. Patent No. 5,143,854; U.S. Patent No. 5,079,600; U.S. Patent No. 4,923,901). Essentially, polypeptides are synthesized on a substrate (microarray) and these polypeptides can be screened with either protein molecules or fragments thereof or nucleic acid molecules in order to screen for either protein molecules or fragments thereof or nucleic acid molecules that specifically bind the target polypeptides (Fodor *et al.*, *Science* 251:767-773 (1991)).

It is understood that one or more of the molecules of the present invention, preferably one or more of the nucleic acid molecules or protein molecules or fragments thereof of the present invention may be utilized in a microarray based method. In a preferred embodiment of the present invention, one or more of the *B. thuringiensis* nucleic acid molecules or protein molecules or fragments thereof of the present invention may be utilized in a microarray based method. A particular preferred microarray embodiment of the present invention is a microarray comprising nucleic acid molecules encoding genes or fragments thereof that are homologues of known genes or nucleic acid molecules that comprise genes or fragments thereof that elicit only limited or no matches to known genes. A further preferred microarray embodiment of the present invention is a microarray comprising nucleic acid molecules having genes or fragments thereof that are homologues of known genes and nucleic acid molecules that comprise genes or fragment thereof that elicit only limited or no matches to known genes.

In a preferred embodiment, the microarray of the present invention comprises at least 10 nucleic acid molecules that specifically hybridize under high stringency to at least 10 nucleic acid molecules encoding *B. thuringiensis* protein or fragments thereof set forth in Table 1. In a more preferred embodiment, the microarray of the present invention comprises at least 100 nucleic acid molecules that specifically hybridize under high stringency to at least 100 nucleic acid molecules that encode a *B. thuringiensis* protein or fragment thereof set forth in Table 1. In an even more preferred embodiment, the

microarray of the present invention comprises at least 1,000 nucleic acid molecules that specifically hybridize under high stringency to at least 1,000 nucleic acid molecules that encode a *B. thuringiensis* protein or fragment thereof set forth in Table 1. In a further even more preferred embodiment, the microarray of the present invention comprises at least 2,500 nucleic acid molecules that specifically hybridize under high stringency to at least 2,500 nucleic acid molecules that encode a *B. thuringiensis* protein or fragment thereof set forth in Table 1. While it is understood that a single nucleic acid molecule may encode more than one protein homologue or fragment thereof, in a preferred embodiment, at least 50%, preferably at least 70%, more preferably at least 80%, even more preferably at least 90% of the nucleic acid molecules that comprise the microarray contain one protein or fragment thereof.

In a preferred embodiment, the microarray of the present invention comprises at least 10 nucleic acid molecules that specifically hybridize under high stringency to at least 10 nucleic acid molecules selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 8283 or fragment thereof or complement of either. In a more preferred embodiment, the microarray of the present invention comprises at least 100 nucleic acid molecules that specifically hybridize under high stringency to at least 100 nucleic acid molecules that encode a *B. thuringiensis* protein or fragment thereof set forth in Table 1. In an even more preferred embodiment, the microarray of the present invention comprises at least 1,000 nucleic acid molecules that specifically hybridize under high stringency to at least 1,000 nucleic acid molecules that encode a *B. thuringiensis* protein or fragment thereof set forth in Table 1. In a further even more preferred embodiment, the microarray of the present invention comprises at least 2,500 nucleic acid molecules that specifically hybridize under high stringency to at least 2,500 nucleic acid molecules that encode a *B. thuringiensis* protein or fragment thereof set forth in Table 1. While it is understood that a single nucleic acid molecule may encode more than one protein homologue or fragment thereof, in a preferred embodiment, at least 50%, preferably at least 70%, more preferably at least 80%, even more preferably at least 90% of the nucleic acid molecules that comprise the microarray contain one protein homologue or fragment thereof.

Nucleic acid molecules of the present invention may be used in site directed mutagenesis. Site-directed mutagenesis may be utilized to modify nucleic acid sequences, particularly as it is a technique that allows one or more of the amino acids encoded by a nucleic acid molecule to be altered (e.g. a threonine to be replaced by a methionine). Three basic methods for site-directed mutagenesis are often employed. These are cassette mutagenesis (Wells *et al.*, *Gene* 34:315-23 (1985)); primer extension (Gilliam *et al.*, *Gene* 12:129-137 (1980)); Zoller and Smith, *Methods Enzymol.* 100:468-500 (1983); and Dalbadie-McFarland *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 79:6409-6413 (1982)) and methods based upon PCR (Scharf *et al.*, *Science* 233:1076-1078 (1986); Higuchi *et al.*, *Nucleic Acids Res.* 16:7351-7367

(1988)). Site-directed mutagenesis approaches are also described in European Patent 0 385 962, European Patent 0 359 472, and PCT Patent Application WO 93/07278.

Site-directed mutagenesis strategies have been applied to plants for both *in vitro* as well as *in vivo* site-directed mutagenesis (Lanz *et al.*, *J. Biol. Chem.* 266:9971-9976 (1991); Kovgan and Zhdanov, *Biotekhnologiya* 5: 148-154, No. 207160n, Chemical Abstracts 110: 225 (1989); Ge *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 86:4037-4041 (1989), Zhu *et al.*, *J. Biol. Chem.* 271:18494-18498 (1996), Chu *et al.*, *Biochemistry* 33:6150-6157 (1994), Small *et al.*, *EMBO J.* 11:1291-1296 (1992), Cho *et al.*, *Mol. Biotechnol.* 8:13-16 (1997), Kita *et al.*, *J. Biol. Chem.* 271:26529-26535 (1996), Jin *et al.*, *Mol. Microbiol.* 7:555-562 (1993), Hatfield and Vierstra, *J. Biol. Chem.* 267:14799-14803 (1992), Zhao *et al.*, *Biochemistry* 31:5093-5099 (1992)).

Any of the nucleic acid molecules of the present invention may either be modified by site-directed mutagenesis or used as, for example, nucleic acid molecules that are used to target other nucleic acid molecules for modification. It is understood that mutants with more than one altered nucleotide can be constructed using techniques that practitioners skilled in the art are familiar with such as isolating restriction fragments and ligating such fragments into an expression vector (*see*, for example, Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press (1989)). In a preferred embodiment of the present invention, one or more of the nucleic acid molecules or fragments thereof of the present invention may be modified by site-directed mutagenesis.

In addition to the above discussed procedures, practitioners are familiar with the standard resource materials which describe specific conditions and procedures for the construction, manipulation and isolation of macromolecules (e.g., DNA molecules, plasmids, etc.), generation of recombinant organisms and the screening and isolating of clones, (*see* for example, Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press (1989); Mailga *et al.*, *Methods in Plant Molecular Biology*, Cold Spring Harbor Press (1995); Birren *et al.*, *Genome Analysis: Analyzing DNA*, 1, Cold Spring Harbor, New York).

#### (a) Computer Media

The nucleotide sequence provided in SEQ ID NO: 1, through SEQ ID NO: 8283 or fragment thereof, or complement thereof, or a nucleotide sequence at least 90% identical, preferably 95%, identical even more preferably 99% or 100% identical to the sequence provided in SEQ ID NO: 1 through SEQ ID NO: 8283 or fragment thereof, or complement thereof, can be "provided" in a variety of mediums to facilitate use. Such a medium can also provide a subset thereof in a form that allows a skilled artisan to examine the sequences.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc, storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently known methods for recording information on computer readable medium to generate media comprising the nucleotide sequence information of the present invention. A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of data processor structuring formats (e.g. text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing one or more of nucleotide sequences of the present invention, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al.*, *J. Mol. Biol.* 215:403-410 (1990)) and BLAZE (Brutlag *et al.*, *Comp. Chem.* 17:203-207 (1993)) search algorithms on a Sybase system can be used to identify open reading frames (ORFs) within the genome that contain homology to ORFs or proteins from other organisms. Such ORFs are protein-encoding fragments within the sequences of the present invention and are useful in producing commercially important proteins such as enzymes used in amino acid biosynthesis, metabolism, transcription, translation, RNA processing, nucleic acid and a protein degradation, protein modification, and DNA replication, restriction, modification, recombination, and repair.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the nucleic acid molecule of the present invention. As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As indicated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory that can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention. As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the sequence of the present invention that match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are available can be used in the computer-based systems of the present invention. Examples of such software include, but are not limited to, MacPattern (EMBL), BLASTIN and BLASTIX (NCBIA). One of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that during searches for commercially important fragments of the nucleic acid molecules of the present invention, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequences the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymatic active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, cis elements, hairpin structures and inducible expression elements (protein binding sequences).

Thus, the present invention further provides an input means for receiving a target sequence, a data storage means for storing the target sequences of the present invention sequence identified using a search means as described above, and an output means for outputting the identified homologous sequences. A variety of structural formats for the input and output means can be used to input and output information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the sequence of the present invention by varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments sequence of the present invention. For example, implementing software which implement the BLAST and BLAZE algorithms (Altschul *et al.*, *J. Mol. Biol.* 215:403-410 (1990)) can be used to identify open frames within the nucleic acid molecules of the present invention. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

The present invention also provides, in another aspect, a method for identifying one or more genes encoding insect inhibitory proteins in the sequences of one or more plasmids of a *Bacillus thuringiensis*, wherein the method of the present invention comprises the steps of: a) isolating and purifying plasmid DNA; b) constructing a DNA library from the isolated and purified plasmid DNA; c) sequencing the DNA library to obtain a set of plasmid DNA sequences; d) comparing the set of DNA sequences with a set of chromosomal DNA sequences, wherein the set of chromosomal DNA sequences comprises sequences depicted in SEQ ID No: 1 through SEQ ID No: 8283; e) identify common sequences, which are identified both in the set of plasmid DNA sequences and in the set of chromosomal DNA sequences; f) subtracting the common sequences from the set of plasmid DNA sequences to obtain a subtracted set of plasmid DNA sequences; g) assembling the subtracted set of DNA sequences to obtain contigs and sequences; h) determining open reading frames in the contigs and sequences; and h) identifying one or more genes encoding insect inhibitory proteins in the sequences of one or more plasmids of said *Bacillus thuringiensis*.

Methods for isolating and purifying plasmid DNA are well known in the art. For example, DNA samples comprising endogenous plasmids from *Bacillus thuringiensis* strains can be isolated and purified using the method of Hansen and Olsen (*J. Bacteriology*, Vol. 135, pp. 227-238 (1978)). Methods for constructing a DNA library are well known in the art.

As used herein, "a contig" refers to a contiguous nucleotide sequence that is assembled from two or more constituent nucleotide sequences that share common or overlapping regions of sequence homology. For example, the nucleotide sequences of two or more nucleic acid fragments can be compared and aligned in order to identify common or overlapping sequences. Where common or overlapping sequences exist between two or more nucleic acid fragments, the sequences (and thus their corresponding nucleic acid fragments) can be assembled into a single contiguous nucleotide sequence.

It is well known to a person skilled in the art that the sequence data from a large scale shotgun sequencing project can be processed and assembled into contigs, which represent a reconstruction of the original chromosomal genome sequence from the cloned fragments. Programs are available in the public domain that can analyze the sequence output and assemble the sequences into larger sequence regions representing contiguous sequences of the target genome. Examples of such programs can be found at, for example, <http://genome.wustl.edu/gsc>, <http://www.sanger.ac.uk>, and <http://www.mbt.washington.edu>. An example of a sequence reading program is Phred (<http://www.mbt.washington.edu>). Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files.

The process of assembling DNA sequence fragments generally involves three phases; the overlap phase, the layout phase and the multi-alignment, or consensus, phase. In the overlap phase, each fragment is compared against every other fragment to determine if they share a common subsequence, an indication that they were potentially sampled from overlapping stretches of the original DNA strand. Pairs of fragments are compared in two ways; 1) with both fragments in the same relative orientation, and 2) with one of the fragments having been reverse complemented. In the layout phase, a series of alternate assemblies or layouts of the fragments based on the pairwise overlaps is generated. A layout specifies the relative locations and orientations of the fragments with respect to each other and is typically visualized as an arrangement of overlapping directed lines, one for each fragment. The general criterion for the layout phase is to produce plausible assemblies of maximum likelihood. In this manner, it can be determined if there is more than one way to put the pieces together and if different solutions appear equally plausible. In such a case, one would return to the lab and obtain additional information to resolve the ambiguity. The multi-alignment, or consensus, phase uses more information than just the pairwise alignments in the layout. The sequences of all the fragments in a layout are simultaneously aligned, giving a final set of contigs representing regions of the target genome. An example of an assembly program is PHRAP, which can be found at <http://chimera.biotech.washington.edu/UWGC/tools/phrap.htm>.

A characteristic feature of a DNA sequence is that it can be compared with other known DNA sequences. Sequence comparisons can be undertaken by determining the similarity of the test or query sequence with sequences in publicly available or propriety databases ("similarity analysis") or by searching for certain motifs ("intrinsic sequence analysis") (e.g. *cis* elements) (Coulson, *Trends in Biotechnology* 12:76-80 (1994); Birren, *et al.*, *Genome Analysis* 1:543-559 (1997)).

Similarity analysis includes database search and alignment. Examples of public databases include the DNA Database of Japan (DDBJ) (<http://www.ddbj.nig.ac.jp/>); Genebank (<http://www.ncbi.nlm.nih.gov/web/Genbank/Index.html>); and the European Molecular Biology Laboratory Nucleic Acid Sequence Database (EMBL) ([http://www.ebi.ac.uk/ebi\\_docs/embl\\_db.html](http://www.ebi.ac.uk/ebi_docs/embl_db.html)).

A number of different search algorithms have been developed, one example of which are the suite of programs referred to as BLAST programs. There are five implementations of BLAST, three designed for nucleotide sequences queries (BLASTN, BLASTX, and TBLASTX) and two designed for protein sequence queries (BLASTP and TBLASTN) (Coulson, *Trends in Biotechnology* 12:76-80 (1994); Birren *et al.*, *Genome Analysis* 1:543-559 (1997)).

BLASTN takes a nucleotide sequence (the query sequence) and its reverse complement and searches them against a nucleotide sequence database. BLASTN was designed for speed, not maximum sensitivity, and may not find distantly related coding sequences. BLASTX takes a nucleotide sequence, translates it in three forward reading frames and three reverse complement reading frames, and then compares the six translations against a protein sequence database. BLASTX is useful for sensitive analysis of preliminary (single-pass) sequence data and is tolerant of sequencing errors (Gish and States, *Nature Genetics* 3:266-272 (1993)). BLASTN and BLASTX may be used in concert for analyzing sequence data (Coulson, *Trends in Biotechnology* 12:76-80 (1994); Birren *et al.*, *Genome Analysis* 1:543-559 (1997)).

Given a nucleotide coding sequence and the predicted protein which may be produced from that sequence, it is often preferable to use the protein as the query sequence to search a database because of the greatly increased sensitivity to detect more subtle relationships. This is due to the larger alphabet of proteins (20 amino acids) compared with the alphabet of nucleic acid sequences (4 bases), where it is far easier to obtain a match by chance. In addition, with nucleotide alignments, only a match (positive score) or a mismatch (negative score) is obtained, but with proteins, the presence of conservative amino acid substitutions can be taken into account. Here, a mismatch may yield a positive score if the non-identical residue has physical/chemical properties similar to the one with which it aligns. Various scoring matrices are used to supply the substitution scores of all possible amino acid alignments. A general purpose scoring system is the BLOSUM62 matrix (Henikoff and Henikoff, *Proteins* 17:49-61 (1993)), which is



currently the default choice for BLAST programs. BLOSUM62 is tailored for alignments of moderately diverged sequences and thus may not yield the best results under all conditions. Altschul, *J. Mol. Biol.* 36:290-300 (1993), uses a combination of three matrices to cover all contingencies. This may improve sensitivity, but at the expense of slower searches. In practice, a single BLOSUM62 matrix is often used but others (PAM40 and PAM250) may be attempted when additional analysis is necessary. Low PAM matrices are directed at detecting very strong but localized sequence similarities, whereas high PAM matrices are directed at detecting long but weak alignments between very distantly related sequences.

Homologues in other organisms are available that can be used for comparative sequence analysis. Multiple alignments are performed to study similarities and differences in a group of related sequences.

CLUSTAL W is a multiple sequence alignment package that performs progressive multiple sequence alignments based on the method of Feng and Doolittle, *J. Mol. Evol.* 25:351-360 (1987). Each pair of sequences is aligned and the distance between each pair is calculated; from this distance matrix, a guide tree is calculated, and all of the sequences are progressively aligned based on this tree. A feature of the program is its sensitivity to the effect of gaps on the alignment; gap penalties are varied to encourage the insertion of gaps in probable loop regions instead of in the middle of structured regions. Users can specify gap penalties, choose between a number of scoring matrices, or supply their own scoring matrix for both the pairwise alignments and the multiple alignments. CLUSTAL W for UNIX and VMS systems is available at: [ftp.ebi.ac.uk](ftp://ftp.ebi.ac.uk). Another program is MACAW (Schuler *et al.*, *Proteins, Struct. Func. Genet.* 9:180-190 (1991), for which both Macintosh and Microsoft Windows versions are available. MACAW uses a graphical interface, provides a choice of several alignment algorithms, and is available by anonymous ftp at: [ncbi.nlm.nih.gov \(directory/pub/macaw\)](ftp://ncbi.nlm.nih.gov/directory/pub/macaw).

Sequence motifs are derived from multiple alignments and can be used to examine individual sequences or an entire database for subtle patterns. With motifs, it is sometimes possible to detect distant relationships that may not be demonstrable based on comparisons of primary sequences alone. Currently, the largest collection of sequence motifs in the world is PROSITE (Bairoch and Bucher, *Nucleic Acid Research* 22:3583-3589 (1994)). PROSITE may be accessed via either the ExPASy server on the World Wide Web or anonymous ftp site. Many commercial sequence analysis packages also provide search programs that use PROSITE data.

A resource for searching protein motifs is the BLOCKS E-mail server developed by S. Henikoff, *Trends Biochem Sci.* 18:267-268 (1993); Henikoff and Henikoff, *Nucleic Acid Research* 19:6565-6572 (1991); Henikoff and Henikoff, *Proteins* 17:49-61 (1993). BLOCKS searches a protein or protein encoding nucleotide sequence against a database of protein motifs or "blocks." Blocks are defined as short, ungapped multiple alignments that represent highly conserved protein patterns. The blocks

themselves are derived from entries in PROSITE as well as other sources. Either a protein or nucleotide query can be submitted to the BLOCKS server; if a nucleotide sequence is submitted, the sequence is translated in all six reading frames and motifs are sought in these conceptual translations. Once the search is completed, the server will return a ranked list of significant matches, along with an alignment of the query sequence to the matched BLOCKS entries.

Conserved protein domains can be represented by two-dimensional matrices, which measure either the frequency or probability of the occurrences of each amino acid residue and deletions or insertions in each position of the domain. This type of model, when used to search against protein databases, is sensitive and usually yields more accurate results than simple motif searches. Two popular implementations of this approach are profile searches (such as GCG program ProfileSearch) and Hidden Markov Models (HMMs) (Krough *et al.*, *J. Mol. Biol.* 235:1501-1531 (1994); Eddy, *Current Opinion in Structural Biology* 6:361-365 (1996)). In both cases, a large number of common protein domains have been converted into profiles, as present in the PROSITE library, or HMM models, as in the Pfam protein domain library (Sonnhammer *et al.*, *Proteins* 28:405-420 (1997)). Pfam contains more than 500 HMM models for enzymes, transcription factors, signal transduction molecules, and structural proteins. Protein databases can be queried with these profiles or HMM models, which will identify proteins containing the domain of interest. For example, HMMSW or HMMFS, two programs in a public domain package called HMMER (Sonnhammer *et al.*, *Proteins* 28:405-420 (1997)) can be used.

PROSITE and BLOCKS represent collected families of protein motifs. Thus, searching these databases entails submitting a single sequence to determine whether or not that sequence is similar to the members of an established family. Programs working in the opposite direction compare a collection of sequences with individual entries in the protein databases. An example of such a program is the Motif Search Tool, or MoST (Tatusov *et al.*, *Proc. Natl. Acad. Sci.* 91:12091-12095 (1994).) On the basis of an aligned set of input sequences, a weight matrix is calculated by using one of four methods (selected by the user); a weight matrix is simply a representation, position by position in an alignment, of how likely a particular amino acid will appear. The calculated weight matrix is then used to search the databases. To increase sensitivity, newly found sequences are added to the original data set, the weight matrix is recalculated, and the search is performed again. This procedure continues until no new sequences are found.

The present invention also further provides, in another aspect, a method for identifying plasmid DNA sequences of a *Bacillus* species, the method comprising the steps of a) identifying a *Bacillus* species strain which does not contain plasmid DNA; b) generating a library of chromosomal genomic DNA from said *Bacillus* species strain which does not contain plasmid DNA; c) obtaining the nucleotide

sequence of said chromosomal genomic DNA; d) identifying a *Bacillus* species strain which contains plasmid DNA; e) generating a library of said *Bacillus* species plasmid DNA; f) obtaining the nucleotide sequence of said plasmid DNA; g) subtracting any common sequences identified in the plasmid DNA which are also identified in the chromosomal genomic DNA; and h) constructing contigs and sequences of said plasmid DNA, wherein said contigs and sequences comprise the plasmid DNA sequence of said *Bacillus* species.

Insect inhibitory protein-encoding nucleic acids of the present invention will find particular uses in the plant protection against insects. For instance, insect-resistant transgenic plants can be generated by introducing the exogenous nucleic acids encoding an insect inhibitory polypeptide or protein or insect inhibitory fragment thereof listed in Table 1. Another example is to engineer transgenic microorganism (bacteria or fungi) to express insect inhibitory polypeptides or proteins of the present invention and then to apply them to the insect food source or allow them to reside in soil surrounding plant roots or on the surface of plant leaves.

The transgenic microorganisms of the present invention may be used to produce *B. thuringiensis* polypeptides or proteins of interest., particularly insect inhibitory polypeptides or proteins. Insect inhibitory polypeptides or proteins or insect inhibitory fragments thereof may be secreted, for example as in bacterial systems, meaning targeted to either the periplasm as for gram negative bacteria or localized to the extracellular space for gram negative or any other type of bacterium, or localized to the intracellular spaces within the cytoplasm. Such compositions may be administered to insects according to methods well known in the art. For example, insect inhibitory polypeptides or proteins of the present invention may be formulated as sprayable compositions or as a bait matrix.

Having now generally described the invention, the same will be more readily understood through reference to the following examples which are provided by way of illustration, and are not intended to be limiting of the present invention, unless specified.

### Example 1

The DNA library designated LIB3237 (Monsanto Company, St. Louis, Missouri, United States of America) was prepared from purified chromosomal DNA of *Bacillus thuringiensis* strain EG10650. *B. thuringiensis* EG10650 is a derivative of strain EG10368 (U.S. Patent No. 5759538; June 2, 1998) that is deficient in neutral protease and alkaline protease activities and contains only one known extrachromosomal plasmid element of 7.5 kb. Deletion mutations in both the alkaline protease and neutral protease genes which were constructed first in strain EG10368 to produce strains EG10654 and EG10624 were combined to produce strain EG10650. Strain EG10650 provides the value described in

US 5,759,538 in which microorganisms containing these modifications exhibit markedly lower proteolytic activity with respect to counterpart microorganisms containing non-disabled protease genes. The result of the reduced proteolytic activity is that the microorganisms expressing insecticidal crystal protein genes and which contain such disabled protease genes produce higher levels of insecticidal crystal proteins and produce crystal proteins having increased stability during storage.

To assure the availability of materials to those interested members of the public upon issuance of a patent on the present application, deposits of the microorganisms listed above were made prior to filing the present application with the ARS Patent Collection, Agricultural Research Service Culture Collection (NRRL), 1815 North University Street, Peoria, Ill. 61604. These microorganism deposits were made under the provisions of the "Budapest Treaty on the International Recognition of the Deposit of Microorganisms for the Purpose of Patent Procedure". EG10654 (NRRL Accession Number NRRL B-21344) and EG10624 (NRRL Accession Number NRRL B-21347) are currently available without restriction. All restrictions on the availability to the public of EG10650 (NRRL Accession Number NRRL B-30217) will be irrevocably removed upon issuance of a United States Patent based on this application. The DNA preparation, chromosomal DNA purification and library construction are described below.

#### DNA Preparation:

*Bacillus thuringiensis* EG10650 was grown under standard conditions in Brain Heart Infusion medium (Difco). Bacterial DNA was prepared in Agarose Plugs, according to the following procedure:

1. Single colony or loop full of bacteria was inoculated in 5 ml of plain Brain Heart Infusion medium and allowed to grow for overnight at 37 ° C;
2. One ml of above culture was inoculated in 20 ml of plain Brain Heart Infusion medium and allowed to grow till 0.6 OD<sub>600</sub>;
3. 180 µg/µl Chloramphenicol was added into the culture and the culture is allowed to grow for one more hour;
4. Cells were harvested by centrifuging at 5,000 g for 10 minutes at 4° C and the harvested cells were washed with 10 ml Solution A and then re-suspended in 1 ml of cold Solution A;
5. The re-suspended cells were brought to room temperature and quickly mixed with equal volume of 2% Seaplaque (FMC Corp.) agarose which was prepared in Solution A and kept at 50° C. The mixture of the cells and agarose was casted as 100 ul plugs in a mold placed on ice;

6. After 1 hour or solidification, the plugs were carefully removed from mold and placed in a tube containing 2 ml of Solution B. The plugs were incubated in a bath at 50°C for 12-16 hours with occasional shaking;
7. Proteinase K (0.5 mg/ml) was added in the tube and incubation was carried out for 16-24 hours; and
- 5 8. Finally the above solution was carefully replaced with 2 ml Solution C with 75 ug/ml Rnase. Incubation with Rnase was carried out for 16-24 hours. The plugs were stored at 4° C in the same solution until they were needed.

Solution A: 1M NaCl, 50mMTris HCl (pH7.5)

Solution B: 50mMTris HCl (pH7.5), 100 mM EDTA, 100 mM NaCl and 2% lysozyme.

- 10 Solution C: 50mMTris HCl (pH7.5), 100 mM EDTA, 100 mM NaCl and 1% SDS.

#### Separation of Chromosomal DNA from Plasmid DNA:

- 15 High molecular weight plasmid DNA was removed from chromosomal DNA for obtaining desired DNA in plugs. Plasmids were removed By CHEF electrophoresis gel. Plugs were subjected to a electrical field of 6 V/cm at an angle of 120 degree in 0.5 X TBE buffer and 1% agarose gel. Initial switch time of 0.47 seconds and final switch time 44.69 seconds with linear ramping was used for 15 hours. After 15 hours plugs were retrieved form gel and were used for isolation of chromosomal DNA by standard methods.

#### Library Construction:

- 20 Purified chromosomal DNA of *Bacillus thuringiensis* strain EG 10650 was sheared to a desired size before size selecting by regular agarose gel electrophoresis. Ends of sheared and size selected DNA were repaired by PFU Polymerase (Stratagene). Repaired DNA was ligated to pSTBlue-1 vector at EcoR V site (Novagen) by using standard reagents and conditions recommended by manufacturer of DNA ligase enzyme (Life Technology). After ligation DNA was introduced by electroporation into
- 25 commercially available *E. coli* DH10B cells (Life Technology). Transformants were plated on LB agar plate with ampicillin for selection of plasmids. Plasmid DNA was prepared by standard procedure for sequencing by T7 and SP6 primers.

#### **Example 2**

- 30 Two basic methods can be used for DNA sequencing, the chain termination method of Sanger *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 74:5463-5467 (1977) and the chemical degradation method of Maxam and Gilbert, *Proc. Natl. Acad. Sci. (U.S.A.)* 74:560-564 (1977). Automation and advances in technology

such as the replacement of radioisotopes with fluorescence-based sequencing have reduced the effort required to sequence DNA (Craxton, *Method*, 2:20-26 (1991); Ju *et al.*, *Proc. Natl. Acad. Sci. (U.S.A.)* 92:4347-4351 (1995); Tabor and Richardson, *Proc. Natl. Acad. Sci. (U.S.A.)* 92:6339-6343 (1995)). Automated sequencers are available from, for example, Pharmacia Biotech, Inc., Piscataway, New Jersey (Pharmacia ALF), LI-COR, Inc., Lincoln, Nebraska (LI-COR 4,000) and Millipore, Bedford, Massachusetts (Millipore BaseStation).

In addition, advances in capillary gel electrophoresis have also reduced the effort required to sequence DNA and such advances provide a rapid high resolution approach for sequencing DNA samples (Swerdlow and Gesteland, *Nucleic Acids Res.* 18:1415-1419 (1990); Smith, *Nature* 349:812-813 (1991); Luckey *et al.*, *Methods Enzymol.* 218:154-172 (1993); Lu *et al.*, *J. Chromatog. A.* 680:497-501 (1994); Carson *et al.*, *Anal. Chem.* 65:3219-3226 (1993); Huang *et al.*, *Anal. Chem.* 64:2149-2154 (1992); Kheterpal *et al.*, *Electrophoresis* 17:1852-1859 (1996); Quesada and Zhang, *Electrophoresis* 17:1841-1851 (1996); Baba, *Yakugaku Zasshi* 117:265-281 (1997)).

A number of sequencing techniques are known in the art, including fluorescence-based sequencing methodologies. These methods have the detection, automation and instrumentation capability necessary for the analysis of large volumes of sequence data. Currently, the 377 DNA Sequencer (Perkin-Elmer Corp., Applied Biosystems Div., Foster City, CA) allows the most rapid electrophoresis and data collection. With these types of automated systems, fluorescent dye-labeled sequence reaction products are detected and data entered directly into the computer, producing a chromatogram that is subsequently viewed, stored, and analyzed using the corresponding software programs. These methods are known to those of skill in the art and have been described and reviewed (Birren *et al.*, *Genome Analysis: Analyzing DNA*, 1, Cold Spring Harbor, New York).

### Example 3

#### Generation and Assembly of *Bacillus thuringiensis* Chromosomal Sequence

This example serves to illustrate the generation of the contigs and singletons listed in the Sequence Listing.

PHRED (phragment editor), which is developed by Phil Green at the University of Washington, was used to call the bases from the sequence trace files and to assign quality scores to the bases. PHRED uses Fourier methods to examine the four base traces in the region surrounding each point in the data set in order to predict a series of evenly spaced predicted locations. That is, it determines where the peaks would be centered if there were no compressions, dropouts, or other factors shifting the peaks from their "true" locations. Next, PHRED examines each trace to find the centers of the actual, or observed peaks

and the areas of these peaks relative to their neighbors. The peaks are detected independently along each of the four traces so many peaks overlap. A dynamic programming algorithm is used to match the observed peaks detected in the second step with the predicted peak locations found in the first step. Default parameters were used in the base calling.

5 After the base calling is completed, sequence preprocessing is performed. Quality assessment and trimming is performed by determining the maximum scoring segment of PHRED quality score > 10. Cloning sequences are removed by utilizing cross\_match (<http://www.mbt.washington.edu>) and searching a database of relevant cloning sequences. Contaminating sequences (E. coli, yeast, vector, linker) are then removed from the dataset by utilizing cross\_match to search a database of contamination sequences.

10 The preprocessed sequences are then assembled into contigs, or groups of overlapping sequences. Contigs are assembled using PHRAP (phragment assembly program), also developed by Green at the University of Washington (<http://www.mbt.washington.edu>) using default assembly parameters. This program takes a file of shotgun sequences and compiles consensus contig sequences. Alignments are influenced by quality scores, based on Green's algorithm. Singletons are the remaining sequences without sufficient overlaps with others after the assembly.

15 A total of 8283 contigs and singletons were obtained. Contig sequences are recognized as those sequences whose designations begin with . Singleton sequences are recognized as those having designations which begin with . All contig and singleton sequences were run through the annotation and gene selection processes as described in Example 4.

#### 20 Example 4

##### Identification of *Bacillus thuringiensis* Genes

25 This example illustrates the identification of genes within the contig and singleton sequences assembled as described in Example 3. The genes and partial genes embedded in such contigs and singletons were identified through a series of informatic analyses.

30 Contigs and singletons are interrogated using AAT-NAP and BLASTP. AAT\_NAP is a program used for constructing a global alignment of a DNA sequence and a protein sequence (Huang, X. *et al.* Genomics 46:37-45 (1997), the entirety of which is herein incorporated by reference). The alignment model of NAP accommodates introns and frameshifts within codons. The scheme for scoring an alignment has several features that allow NAP to identify the exact locations of introns. A nucleotide insertion gap of length  $\leq k$  is given a linear penalty, and a nucleotide insertion gap of length  $> k$  is penalized as a gap of length  $k$ , where the value for  $k$  is the default value. The NAP program reports the starting and ending coordinates of predicted genes. The input to the NAP program includes the query

sequence, the protein database and a coordinate file produced by AAT\_EXT (an adapter between a database search program and a sequence alignment program) from the output of AAT\_DPS ( a program computing high-scoring chains of segment pairs between a query DNA sequence and the public non-redundant protein database from NCBI) The NAP program scans the protein database and finds the protein sequence for each coordinate record. Then for each coordinate record, NAP locates the query region, extends the region in both directions by a certain number of bases, and computes an alignment of the extended region and the protein sequence. NAP corrects frameshifts in the query sequence.

BLASTP is used to validate the amino acid sequences and hits reported by the AAT\_NAP program and to assign BLAST scores and p values to each sequence/hit pair. The AAT\_NAP generated amino acid sequences are compared with the public non-redundant protein database (nr.aa from NCBI) using the default BLASTP parameters except that the V parameter is set to 1000000 (to report up to 1000000 hits that exceed the BLASTP default report cutoff) If the hit reported by AAT\_NAP for a particular amino acid sequence is not reported by BLASTP, that particular amino acid sequence is removed. Protein encoding regions in the *Bacillus thuringiensis* nucleic acid molecules of the present invention are listed in Table 1.

Coding sequences identified in Table 1 encode many useful *B. thuringiensis* polypeptides or proteins or fragments thereof, including but not limited to insect inhibitory polypeptides or proteins, polypeptides or proteins capable of conferring antibiotic resistance, cytotoxin proteins which may be used as microbial inhibitory proteins including bactericidal, bacteriostatic, fungicidal, and fungistatic polypeptides or proteins, polyketide synthases, polypeptides or proteins capable of conferring resistance to heavy metals or other chemicals, transposons and mobile genetic elements and their corresponding transposases, excisases, integrases, and invertases, phage and phage particle proteins, transcription regulatory proteins, translation regulatory proteins, and other useful proteins homologous to proteins.